Universidad Zaragoza
1542

Trabajo Fin de Máster

Generación automática de metadatos geográficos de páginas Web

Autor

Bernardo José Borjas Borjas

Director

Francisco Javier Zarazaga Soria

Escuela de Ingeniería y Arquitectura

2011

Agradecimientos

Este trabajo ha sido posible gracias a la oportunidad presentada por el Grupo de Sistemas de Información Avanzados (IAAA) del Departamento de Ingeniería de Sistemas e Informática de la Universidad de Zaragoza, especialmente al Doctor Francisco Javier Zarazaga-Soria, por su apoyo incondicional, ético y profesional.

Al grupo de colaboradores, en particular a la Doctora Aneta Jadwiga Florczyk, por su don de compañerismo, apoyo permanente y trabajo colaborativo.

A Rosa Mary, por su amor, paciencia y comprensión.

A José Alberto, por iluminarme en todo momento con su constante e inocente sonrisa.

Resumen

En la Web se están poniendo en línea muchos recursos e irá aumentando, por ejemplo, teniendo en cuenta la recientemente declarada transparencia y liberación de datos públicos por parte de los gobiernos. Según diversos estudios, el 80% de toda la información almacenada en soporte electrónico por las Administraciones Públicas está hasta ahora relacionada con alguna localización geográfica (información georreferenciada) o es susceptible de estarlo. En realidad, cualquier tipo de información tiene aspecto espacial y suele contener información geográfica de forma implícita o explicita. Eso se refiere también a los contenidos publicados en la Web.

Por otro lado, desde que surge el concepto de Infraestructura de datos Espaciales (IDE) en 1994 se ha llevado a cabo un gran avance en el desarrollo de conceptos, modelos y arquitecturas que han posibilitado la creación de una sólida base. Esta base ha permitido la puesta en funcionamiento de un número cada vez más importante de IDEs a nivel local, regional, nacional y transnacional. Los georrecursos Web que no forman parte de una IDE pueden ser una fuente importante de información, o por lo menos, una fuente complementaria, pero su incorporación requiere de un gran esfuerzo para la creación de su descripción, que es causado por la diferencia que existe entre la aproximación al descubrimiento de recursos en la Web y dentro de una IDE.

Los metadatos estandarizados son un arma poderosa porque permiten descubrir y seleccionar los recursos digitales relevantes de una forma fácil y rápida. La creación de metadatos es un proceso costoso y, en la práctica, en la Web continuamente se crea mucho contenido, pero sin describirlos. Un modo de resolver este problema es conseguir que no sea necesario que los publicadores de recursos se tengan que preocupar por la creación de los metadatos. Esto nos lleva a la necesidad de generar metadatos de forma automatizada.

Este Trabajo de Fin de Máster se dedicó al desarrollo de una arquitectura para la generación automática de metadatos geográficos para recursos de Web, con aspecto extensible y flexibilidad para la adición de nuevas características. Para el estudio de un caso de uso se desarrolló un prototipo que se empleó para la generación de registros OGC CSW que describen a los recursos Web. El primer experimento realizado para la validación del prototipo, sobre una muestra representativa de páginas Web principales de geoportales (de ámbito global, regional, nacional y local), ha demostrado que el principal problema era la generación de información sobre la extensión geográfica, ya que las páginas Web no suelen contener metadatos geográficos específicos. Por esta razón, el sistema se complementó con el uso de una herramienta NER que aplica algoritmos NLP para la extracción de nombres de lugares del texto y el desarrollo de un componente para la estimación de la extensión geográfica (*Bounding Box*) que contempla los nombres geográficos encontrados dentro de los diferentes elementos de una página Web. Los resultados del segundo experimento pueden indicar que usando una heurística muy simple (basada en la frecuencia de nombres geográficos y la agrupación según la pertenencia a una unidad de organización territorial) se puede estimar la extensión geográfica, con un nivel satisfactorio, en casi un 70%.

Tabla de contenidos

Índice de tablas y figuras

Tablas:

Figuras:

Capítulo 1. Introducción

1.1. Contexto

Dentro del Programa Oficial de Posgrado en Ingeniería Informática, adaptado al Espacio Europeo de Educación Superior (EEES), que conduce a la obtención del Máster en Ingeniería de Sistemas e Informática, se requiere la realización de un trabajo o proyecto de iniciación a la investigación o innovación tecnológica.

Este Trabajo de Fin de Máster (TFM) se ha realizado en el entorno de desarrollo del grupo de Sistemas de Información Avanzados (IAAA)[1], adscrito al Instituto de Investigación en Ingeniería de Aragón (I3A)[2] de la Universidad de Zaragoza. La actividad de investigación del grupo está enfocada en las tecnologías de software abierto, distribuido e interoperable, principalmente mediante servicios Web y para sistemas de información geoespacial, abarcando áreas como Sistemas de Información Geográfica (SIG), Teledetección, Servicios Basados en la Localización (SBL) y, con una atención especial a las Infraestructuras de Datos Espaciales (IDEs). Concretamente, la labor de investigación del grupo aborda problemáticas vinculadas a las ontologías y metadatos; interoperabilidad, composición y encadenamiento de servicios; visualización inteligente de información geográfica, metadatos, procesamiento semántico y recuperación inteligente de la información (indexación, interrogación y recuperación); métodos y procesos para la creación de información y el modelado de contenidos heterogéneos; y, finalmente, un aspecto que se considera fundamental para la realimentación técnica como lo es la definición, creación y puesta en funcionamiento de nuevos servicios y aplicaciones, integrando utilidades de geoprocesamiento en sistemas de información tradicionales como: servicios distribuidos de catálogo de información geográfica, servidores de mapas, *features, coverages, gazetteers*, geocodificadores, *geoparsers*, etc.

En la Web se están poniendo en línea muchos recursos e irá aumentando, por ejemplo, teniendo en cuenta la recientemente declarada transparencia y liberación de datos públicos por parte de los gobiernos [1]. Según diversos estudios [2], el 80% de toda la información almacenada en soporte electrónico por las Administraciones Públicas está hasta ahora relacionada con alguna localización geográfica (información georreferenciada) o es susceptible de estarlo. En realidad, cualquier tipo de información tiene aspecto espacial y suele contener información geográfica de forma implícita o explicita. Eso se refiere también a los contenidos publicados en la Web. Una página Web, por ejemplo, puede estar vinculada a una o varias localizaciones geográficas según:

- su IP que es georreferenciable,
- en el URL se podrían identificar nombres de lugares,
- los nombres geográficos o coordenadas explícitas que pueden aparecer en el contenido de la página (p. ej. metadatos, texto plano), o
- se puede inferir a base de elementos relacionados (p. ej. vía enlaces).

Gran parte del contenido publicado en la Web son recursos de base geográfica, que pertenecen a la Web geoespacial, una de las capas entrelazadas de la Web actual. Principalmente, los georecursos son aquellos recursos que siguen los estándares, las especificaciones y/o las recomendaciones de los organismos internacionales dedicados a la información geográfica, como

[1] http://iaaa.cps.unizar.es/
[2] http://i3a.unizar.es/

la *Open Geospatial Consortium, Inc.*[3] (OGC), la *International Organisation for Standarisation*[4] (ISO), *World Wide Web Consortium*[5](W3C), o los recursos con formato procedente del sector privado pero usado popularmente en la Web (p. ej. *Shapefile* [3]). En general, los georrecursos son:

- ficheros de texto con datos espaciales en formato geográfico, por ejemplo KML [4], GML [5];
- recursos con publicación de datos espaciales (p. ej. servicios de mapas);
- ficheros de texto de formatos dedicados a la descripción de otros recursos geográficos (p. ej. Dublín Core (DC)) [6], normas ISO 19115/ISO 19139 [7,8], *OGC Web Service Commons* [9]) o donde la información geográfica es información contextual (p. ej. GeoRSS[6] o RDF con vocabulario geográfico como *GeoNames Ontology*[7]);
- imágenes o vídeos con información espacial incrustada (p. ej. geotiff [10]) o asociada (p. ej. etiqueta), que puede ser explícita (p. ej. coordenadas con sistema de referencia) o implícita (por ejemplo, nombre del lugar).

Parte de los recursos de la Web geoespacial forma parte de la Infraestructura de Datos Espaciales (IDE), por ejemplo, los servicios Web publicados en portal IDEE[8] por el Instituto Geográfico Nacional de España[9]. En [11] se define IDE como *la colección base de referencia de tecnologías, políticas y acuerdos institucionales que faciliten la disponibilidad y el acceso a la información espacial*, la cual *proporciona una base para el descubrimiento de datos espaciales, la evaluación y la aplicación para los usuarios y los proveedores en todos los niveles de gobierno, el sector comercial, el sector sin fines de lucro, las instituciones académicas y de los ciudadanos en general.* Un ejemplo de una IDE nacional es la Infraestructura de Datos Espaciales de España, IDEE, y la Infraestructura de Información Espacial en la Unión Europea, INSPIRE[10], es una iniciativa internacional. Desde el nacimiento del concepto de IDE (1994) se ha llevado a cabo un gran avance en el desarrollo de conceptos, modelos y arquitecturas que han posibilitado la creación de una base sólida. Esta base ha permitido la puesta en funcionamiento de un número cada vez más importante de IDEs a nivel local, regional, nacional y transnacional. No obstante, basta con echar un rápido vistazo para observar que el eje central sobre el cual se sustenta la mayor parte de estas iniciativas lo constituye la información geográfica más tradicional (mapas, coberturas, modelos digitales del terreno, etc.). Para ello se ha venido utilizado, como soporte de caracterización de esta información, la norma ISO. Más recientemente, se ha abierto la puerta al trabajo con contenidos heterogéneos de la mano de los estándares con un nivel de abstracción superior tales como DC y, a partir de aquí, establecer mecanismos de mapeo sobre esquemas y estándares que posibiliten la descripción de otros tipos de elementos, con un reaprovechamiento de las infraestructuras de soporte al almacenamiento y de los servicios de búsqueda ya desarrollados.

En el ámbito de la Web en general hay que destacar que los estándares abiertos y las pautas son el objetivo principal de la W3C, una organización sin fines de lucro, fundada en 1994, donde sus miembros, procedentes de distintas partes del mundo, tienen la misión de guiar a la Web hacia su máximo potencial a través del desarrollo de protocolos y directrices que aseguren el crecimiento de la Web a largo plazo. Para llevar a cabo esta labor, el W3C une a diversos agentes sociales, bajo un proceso claro y efectivo basado en el consenso para desarrollar estándares de alta calidad que

[3] http://www.opengeospatial.org/
[4] http://www.iso.org/
[5] http://www.w3.org/
[6] http://www.georss.org/
[7] http://www.geonames.org/ontology/documentation.html
[8] http://www.idee.es/
[9] http://www.ign.es/
[10] http://inspire.jrc.ec.europa.eu/

toman como base las contribuciones aportadas por sus Miembros[11], su Equipo[12] y la comunidad en general.

1.2. Motivación

Los georrecursos Web que no forman parte de una IDE pueden ser una fuente importante de información, o por lo menos, una fuente complementaria. Por un lado existe la información generada por los especialistas en el campo de la información geográfica pero que no está incorporada dentro de una IDE (es decir, no publicada para su descubrimiento), y por otro lado, hay información geográfica generada por las comunidades de usuarios de la Web, denominada como "neo geography" [12], "naïve geography" [13] o "volunteered geographic information" (VGI) [14]. Los recientes trabajos de investigación subrayan la importancia de aprovechar VGI como una fuente más de información para una IDE [15,16].

Independientemente de la procedencia, los recursos Web de base geográfica que no están dentro de una IDE podrían enriquecerla, pero su incorporación requiere de un gran esfuerzo para la creación de su descripción, que es causado por la diferencia que existe entre la aproximación al descubrimiento de recursos en la Web y dentro de una IDE. La búsqueda general de los recursos en la Web se basa principalmente en el uso de motores de búsqueda, cuyos contenidos han sido creados mediante el uso de los *crawlers* (agentes 'robot'), a base del análisis automático del contenido de los recursos y las relaciones entre ellos [17]. También existen las bibliotecas digitales en la Web creadas y mantenidas por los seres humanos, pero suelen especializarse en campos concretos como, por ejemplo, la medicina [18]. El descubrimiento de los recursos en el contexto de IDE se basa en el paradigma de las bibliotecas digitales, donde los metadatos (principalmente DC e ISO 19115) contienen la descripción del recurso y están recogidos en catálogos para su búsqueda y recuperación [19].

Los metadatos estandarizados son un arma poderosa porque permiten descubrir y seleccionar los recursos digitales relevantes de una forma fácil y rápida. El estándar de metadatos *Dublin Core* es un conjunto de elementos, simple pero eficaz, que permite describir una amplia gama de recursos en la red, donde cada uno de estos elementos es opcional y puede repetirse. La creación de metadatos es un proceso costoso y, en la práctica, en la Web continuamente se crea mucho contenido, pero sin describirlos. Un modo de resolver este problema es conseguir que no sea necesario que los publicadores de recursos se tengan que preocupar por la creación de los metadatos. Esto nos lleva a la necesidad de generar metadatos de forma automatizada.

1.3. Objetivos

En este Trabajo Fin de Máster se pretende avanzar metodológicamente y tecnológicamente en los procesos de generación de metadatos geográficos que caractericen páginas Web. Para ello, se establecen los siguientes objetivos, tomando en cuenta los objetivos implícitos en el desarrollo de los mismos:

a) En primer lugar, revisar el estado del arte e identificar los trabajos más relevantes de relacionados con la extracción y generación de metadatos a partir de páginas Web

b) Evaluar las propuestas de mayor éxito y plantear una propuesta de arquitectura de un sistema dedicado a la caracterización automática de recursos en Internet de páginas Web

[11] http://www.w3.org/Consortium/Member/List
[12] http://www.w3.org/People/

clásicas, ajustada a los requerimientos de una Infraestructura de Datos Espaciales (IDE), cuyo modelo de metadatos corresponda al modelo de recuperación utilizado por los catálogos de metadatos de recursos dentro de la misma.

c) El reto es un sistema extensible que sea factible de incorporar el soporte a diferentes tipos de recursos Web pero también de posibilitar al usuario para modificar la lógica de generación de los metadatos "al vuelo".

d) Validar la propuesta del sistema de extracción automática de las propiedades usadas comúnmente por los servicios de catálogo de Web de la OGC (OGC CSW) para describir un recurso, sobre una muestra representativa de páginas Web de proveedores de IDE a nivel global, regional, nacional y local.

1.4. Estructura

Este documento se ha dividido en cinco secciones principales:

- Este Capítulo 1, de introducción, permite contextualizar el ámbito profesional y referenciar los entornos sociales involucrados, además, de los objetivos del desarrollo de la temática presentada.
- El Capítulo 2 recoge el estado del arte donde se describen las aportaciones al conocimiento producidas por las investigaciones y las técnicas de mayor relevancia tecnológica.
- El Capítulo 3 describe el diseño y arquitectura del sistema propuesto. Se presenta una arquitectura de caracterización automática de recursos en Internet de páginas Web clásicas que interactúa con diferentes componentes de servicios interoperables.
- El Capítulo 4 describe la implementación del caso de uso utilizado como prototipo de la arquitectura del sistema propuesto y los experimentos efectuados.
- El Capítulo 5 describe algunas conclusiones provenientes de la experiencia de desarrollo y ejecución del prototipo planteado. Adicionalmente, también se plantean algunas líneas futuras de trabajo.

Por último, este documento se completa con un capítulo de bibliografía y unos anexos donde se recogen: acrónimos, una descripción de las herramientas y tecnologías utilizadas, la relación de los geoportales utilizados como muestra representativa en los experimentos y los detalles de diseño del prototipo.

Capítulo 2. Estado del Arte

2.1. Generación de Metadatos

2.1.1. General

La generación de metadatos es el acto de creación o producción de metadatos. La generación de metadatos de buena calidad de forma eficaz es esencial para organizar y facilitar la accesibilidad al creciente número de recursos de información, ricos en contenido, que se disponen cada día. El éxito de las bibliotecas o catálogos digitales, la preservación de la información y el sustento de la interoperabilidad - promovido por las diferentes iniciativas de acceso abierto a la información - y la evolución de la Web Semántica se basan todos en la generación eficiente de los metadatos.

En la figura 1 se muestra, de forma esquemática, a la generación de metadatos que involucra procesos, personas y herramientas [20].

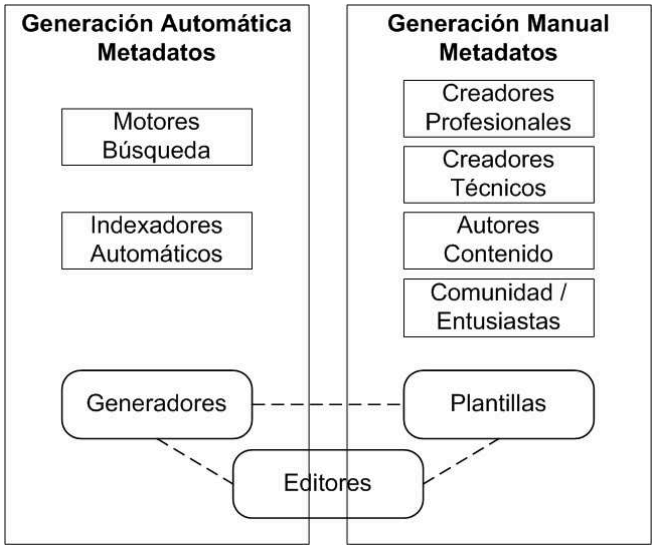

Figura 1: Generación de metadatos

La generación de metadatos manual tiene lugar cuando una persona crea los metadatos. La calidad de los metadatos generados manualmente está determinada por la adhesión, de forma semántica y sintáctica, a un esquema de especificación de metadatos. La prevalencia de este tipo de metadatos puede atribuirse, en parte, al hecho de que la producción de metadatos a través del intelecto humano es a menudo superior a los metadatos producidos de forma automática.

Las personas involucradas en la generación de metadatos se pueden agrupar de acuerdo a la siguiente clasificación:

- creadores profesionales de metadatos (con entrenamiento formal y competencia en el uso estándares),
- creadores técnicos de metadatos (con entrenamiento básico para trabajar con esquemas simples y la ejecución de procesos rutinarios),
- creadores o autores de contenido intelectual y,
- las comunidades de interés o los entusiastas en temas específicos.

Las plantillas *(templates*) de metadatos son hojas de referencia básica que proporcionan un resumen global de los elementos de un esquema de especificación de los metadatos. La plantilla, tanto en formato impreso y electrónico, ha sido una herramienta predominante en la generación manual de metadatos, debido a que son fáciles de producir y de mantener.

La generación automática de metadatos depende del procesamiento de una máquina. En el contexto de la Web, son conocidas las herramientas de indexación o clasificación automática, que se centran en la temática del contenido del recurso, y los motores de búsqueda comerciales. Los motores de búsqueda practican la generación automática de metadatos bajo dos situaciones. En primer lugar, antes de la búsqueda por parte de un usuario, los *crawlers* recorren la Web y almacenan los metadatos en la base de datos del motor de búsqueda. La primera ejecución de la búsqueda de un usuario se realiza contra este almacén de metadatos. En segundo lugar, en el momento de la búsqueda los metadatos se producen de forma automática y dinámica mediante la ejecución de un algoritmo de búsqueda y recuperación contra el almacén global de recursos de la Web. La segunda situación se produce cuando la consulta del usuario no coincide con los metadatos almacenados en la base de datos del motor de búsqueda.

Los generadores de metadatos son mecanismos que soportan la producción automática de metadatos. Dentro del contexto de la Web, los generadores requieren primero la presentación de un localizador único del recurso, URL, localizador persistente uniforme para recurso, PURL u otra dirección Web con el fin de localizar el objeto. Un algoritmo se utiliza para revisar minuciosamente el contenido un objeto, incluyendo su código fuente y, automáticamente, capturar y asignar los metadatos.

La generación de metadatos de forma semiautomática o híbrida permite editar y completar los metadatos que se generan de forma automática. Los editores son herramientas similares a las plantillas ya que requieren la intervención humana, pero pueden ser más sofisticados al proporcionar el acceso directo o automatizado a los estándares y a la documentación (esquemas de especificación de metadatos, guías de contenido, tesauros, listas de clasificación, etc.) que guían proceso de creación de los metadatos. Una característica relevante de este tipo de herramientas es que permiten lidiar con el carácter experimental e impredecible de los generadores.

2.1.2. IDE

En el contexto de la IDE, los metadatos se suelen crear de forma manual por expertos, una labor que es muy costosa (en términos de tiempo y esfuerzo), pero que a su vez garantiza una alta calidad, o por los propios productores de datos espaciales [21]. Hay algunos trabajos relacionados con la generación automática de metadatos geográficos dedicados a la IDE, basados fundamentalmente en la indexación, abstracción y clasificación de datos de contenido espacial. Por ejemplo, [21] propone una plataforma, basada en conceptos de etiquetado e indexación social (*folksonomy*), para la creación, actualización y enriquecimiento de los metadatos espaciales de forma automática. Hay muchas herramientas que permiten a un usuario generar de forma automática los metadatos para una variedad de recursos geoespaciales. Por ejemplo, CatMDEdit[13] es un editor de metadatos donde se pueden generar los metadatos de diferentes formatos de ficheros de datos espaciales, de una serie espacial y de servicios Web OWS (a través del análisis de la respuesta del operador OWS *getCapabilities*). Las soluciones propuestas por la comunidad de la IDE son apropiados para los recursos de la Web geoespacial. Sin embargo, no pueden aplicarse con éxito a los georecursos Web, tales como la página Web de un geoportal. Por lo tanto, en este trabajo se ha examinado el uso de los metadatos en las páginas Web y se han estudiado algunos

[13] http://catmdedit.sourceforge.net/

de los principales trabajos de investigación relacionados sobre el tema de la comunidad de la Web. Los enfoques descritos y las herramientas revisadas se han considerado para el diseño de la arquitectura propuesta.

2.1.3. Web

Existe mucho trabajo hecho en el campo de la investigación sobre la creación y el mantenimiento de metadatos que proponen diversos modelos especializados en diferentes tipos de recursos digitales en la Web [22, 23, 24]. [25] muestra que los no profesionales pueden ser tan buenos como los profesionales en la creación de metadatos para los recursos Web, pero los publicadores de contenidos Web no prestan la suficiente atención para asegurarse que la descripción esté adecuada a los recursos o la falsean de forma intencionada [26]. Los metadatos que se describen en las páginas HTML podrían no ser fiables, ya que al ser usados éstos por los *crawlers* para la indexación de los recursos Web, los publicadores de contenido utilizan estos metadatos con el propósito de ganar visibilidad en los motores de búsqueda.

El recurso más popular en la Web es una página HTML y su especificación (la recomendación del W3C) contempla una sección declarativa de cabecera, que contiene información acerca del documento que no es considerado contenido del propio documento. En la cabecera los autores pueden especificar metadatos a través del elemento `META`, en el cual se especifica un par o dupla de propiedad-valor. La tabla 1 recoge los atributos de elemento `META` según W3C.

Atributo	Descripción	O/I
Name	Identifica el nombre de la propiedad	I
Content	Especifica el valor de la propiedad	I
Schema	Identifica esquema para ser usado para interpretar el valor de la propiedad	O
http-equiv	Obtiene información de cabecera en respuesta HTTP (en lugar del *name*)	I

Tabla 1: El elemento `META` (O/I – Opcional/Imperativo)

Esta especificación no enumera los valores legales para el atributo *name*. Los convenios utilizados para definir el valor de los atributos de los elementos `META` se pueden describir, opcionalmente, en lo que la especificación de HTML denomina perfil de metadatos. Un perfil de metadatos se describe mediante un identificador uniforme de recurso (URI) y su uso se divulga en el valor del atributo *profile* del elemento `HEAD`. Por ejemplo, la comunidad *Dublin Core Metadata Iniciative* (DCMI) recomienda un perfil[14] de metadatos mediante el cual se describe como un conjunto de descripción de metadatos *Dublin Core* puede ser codificado usando elementos y atributos HTML. La tabla 2 recoge los elementos `META` usados en HTML, aplicando alineamiento entre los elementos básicos de la recomendación DCMI , la propuesta de W3C y WHATWG [27,28] y algunos elementos `META` usados frecuentemente en las páginas Web (de acuerdo al sitio comercial Metatags.org[15]). En la tabla 3 se muestran los metadatos que se usan popularmente por los elementos `META` para indicar extensión geográfica vinculada al recurso.

[14] http://dublincore.org/documents/2008/08/04/dc-html/
[15] http://www.metatags.org/

DCMI	W3C, WHATWG	Metatags.org	Descripción
DC.contributor			Entidad responsable contribuciones
DC.coverage	geographic-coverage geo.country geo.placename geo.position geo.region icbm		Característica espacial
DC.creator	author	author	Responsable de creación contenido
DC.date	created	creation_date	Fecha de creación o disponibilidad
DC.description	description	description	Descripción textual del contenido
DC.format		content-type (http-equiv)	Formato, medio físico o dimensiones
DC.identifier		identifier-URL	Referencia no ambigua del recurso
DC.language		language	Idioma del recurso
DC.publisher	publisher	webauthor	Entidad responsable de publicación
DC.relation			Identificador recurso relacionado
DC.rights	rights	copyright	Gestión de derechos del contenido
DC.source			Recurso relacionado
DC.subject	keywords	keywords	Palabras claves/frases del contenido
DC.title	application-name		Nombre dado al recurso
DC.type		resource type (http-equiv)	Naturaleza o género del recurso
	creator		Creador del contenido
	audience		Audiencia
	datetime-*		Característica temporal contenido
	rights-standard		Gestión de derechos estandarizados
	Subj-*		Clasificación por temas contenido
		abstract	Resumen
		contact	Contacto
		distribution	Nivel o grado de distribución

Tabla 2: Comparativa de elementos META

Elemento META	Formato	Nota	Ejemplo
ICBM[16]	latitud, longitud	Sistema de coordenadas cartográficas WGS 84	*"51.665471, 6.880063"*
geo.position (geotags[17])	latitud;longitud	Sistema de coordenadas cartográficas WGS 84	*"51.665471;6.880063"*
geo.placename (geotags)	Nombre lugar		*"Steinbergweg, 46514 Schermbeck, Germany"*
geo.region (geotags)	Código subdivisión País	Código ISO 3166-2[18]	*"DE-Nordrhein-Westfalen"*
DC.coverage[.x/y/z/placeName/longitude/latitude]	x/y/altura/Nombre de lugar/longitud/ latitud	Puede estar especificado en x, y, z o nombre de lugar [19]. El sistema de coordenadas debe estar definido por el atributo *scheme*. En caso de longitud latitude es WGS 84.	*"World", "51.665471"*
geographic-coverage	Place-classe, lower-case / código	Definición de región, según WHATWG	*"city, Sao Paulo, Sao Paulo, Brazil"*

Tabla 3: Elementos META geográficos

La generación automática de metadatos está todavía en su infancia, pero algunos enfoques han surgido que incluyen la recolección de metadatos (*harvesting*), la extracción de contenido, la clasificación o la indexación automática, la minería de texto y de datos, el etiquetado social y la generación de metadatos relacionados con la información contextual asociada o recursos relacionados [29]. Las aproximaciones como *DC-dot* generan metadatos *Dublin Core* de una página Web a base de la recolección de elementos de metadatos *(title, keywords, description y type*) desde la cabecera del documento fuente. En ausencia del elemento META, el metadato *keywords* se extrae desde el cuerpo de la página, analizando los conceptos de los hipervínculos (anclajes) y la codificación de presentación (la fuente en negrita, el tamaño de fuente, etc.).

Más extensible y sofisticada es una solución como *Data Fountains* que permite tanto descubrir como describir recursos de Internet. De forma alternativa, puede rastrear y generar metadatos para recursos de Internet sobre un tema particular. O puede profundizar y seguir enlaces desde un URL de inicio. *Data Fountains* genera el rango completo de los elementos simples de *Dublin Core*. El método de generación de metadatos varía desde la simple recolección desde los elementos META en un documento HTML hasta la combinación de esta recolección con un método muy sofisticado para generar palabras claves, o de forma más precisa, frases claves y resúmenes. Una frase clave consiste en una secuencia corta de palabras, en general, un segmento breve de frase, que encapsula un tema clave del documento de origen. Un algoritmo de procesammiento de lenguaje natural, *Natural Language Processing* (NLP), *phraseRate* [30] analiza la estructura del documento HTML y usa este conocimiento como base para la identificación y clasificación de las frases claves candidatas. Utiliza una matriz de indicadores, que incluye la estructura anidada del documento, la estructura gramatical y sintáctica de las oraciones; la frecuencia relativa de cada palabra, su posición dentro de una frase clave candidata, y la frecuencia de dichas frases claves dentro del documento fuente, considerado como un todo; y el peso relativo que puede darse, respectivamente, a las frases claves extraídas desde el cuerpo y a las palabras claves de los elementos META. En [31, 32] se hace referencia a un servicio comercial australiano llamado *Klarity* (actualmente no accesible) que permite generar de forma automática metadatos para cinco elementos de metadatos *(identifier, title, concepts, keywords y description*). Estos

[16] http://geourl.org/add.html
[17] http://geotags.com
[18] http://www.iso.org/iso/country_codes/background_on_iso_3166/iso_3166-2.htm
[19] http://alexandria.ucsb.edu/public-documents/metadata/dc_coverage.html

metadatos son convertidos a etiquetas `META` o al formato XML-RDF. El elemento *'concepts'* de *Klarity* es un concepto unificador de palabras claves y funciones, que se crea a partir de un algoritmo basado en frecuencias de términos que se comparan con un vocabulario subyacente. Esta herramienta dispone también de un editor para introducir de forma manual metadatos adicionales respondiendo a una serie de preguntas.

Apache Tika es un conjunto de herramientas, basado en *Java*, que ofrece una interfaz genérica para detectar, analizar y extraer metadatos y contenido de texto estructurado a partir de diversos tipos de formatos de documentos de acuerdo a los tipos MIME [20]. La lógica de extracción de contenido no se encuentra dentro del propio *Tika*, sino que utiliza o delega a los analizadores sintácticos (*Parsers*) ya existentes, ocultando su complejidad [33]. Por ejemplo, en el caso del código HTML, se utiliza el analizador *TagSoup*[21] .

2.2. Áreas Relacionadas

2.2.1. Herramientas NER

El reconocimiento de nombre de entidades o *Named Entity Recognition* (NER), abarca el procesamiento de un texto y la identificación de ciertas ocurrencias de palabras o expresiones que pertenecen a determinadas categorías de nombre de entidades. Los sistemas de reconocimiento de nombres de entidades son una herramienta importante de preprocesamiento para tareas como la extracción de información, la recuperación de la información y otras aplicaciones de procesamiento de textos [34].

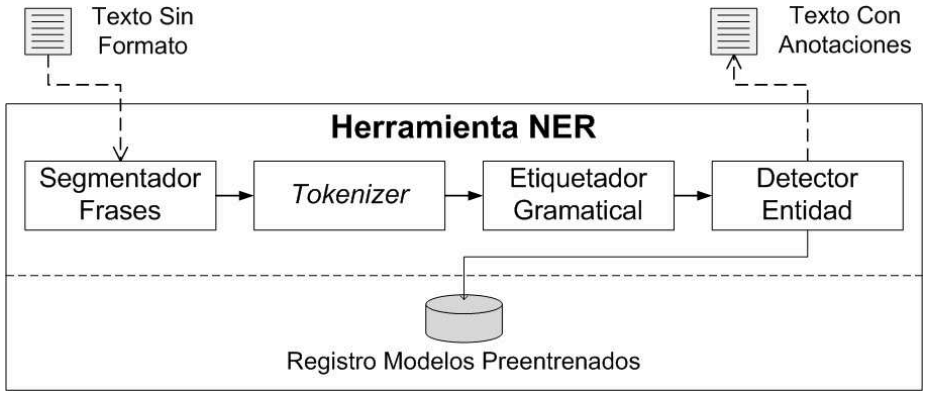

Figura 2: Arquitectura de un sistema NER simple

La figura 2 muestra la arquitectura para un sistema NER simple. El texto sin formato del documento se divide en frases usando un segmentador de frases, y cada frase se subdivide a su vez en palabras usando un *Tokenizer*. A continuación, mediante el etiquetador gramatical (conocido como *Part-of-speech tagging* - POST), se etiqueta cada una de las palabras de una frase con su categoría gramatical, que será de gran utilidad en el próximo paso con el uso del detector de nombre de entidad, el cual trata de localizar y clasificar (mediante anotaciones en el texto original) los elementos atómicos en el texto en categorías predefinidas, como por ejemplo los nombres de personas, organizaciones, lugares, expresiones de los tiempos, cantidades, valores monetarios, porcentajes, etc. Para ser capaz de detectar el nombre de una entidad, el detector de

[20] http://www.ietf.org/rfc/rfc3851.txt
[21] http://home.ccil.org/~cowan/tagsoup/

nombres de entidades requiere de un modelo que suele depender del idioma y la entidad para la cual se entrenó de forma previa.

Entre los ejemplos de implementación de herramientas NER, basados en *Java*, se pueden citar las siguientes:

- *Stanford* NER (*CRFClassifier*)[22],
- *Apache OpenNLP* (*Name Finder*)[23]
- *Illinois* NER *Tagger*[24].

2.2.2. Georreferenciación

El reconocimiento de nombres de entidades (NER) involucra la identificación de nombres propios dentro del texto. La identificación de nombres geográficos por si sola no es suficiente para asignar la extensión geográfica a un recurso Web. Para ello es necesario un proceso complementario: la georreferenciación.

La georreferenciación o codificación geográfica se refiere al posicionamiento con el que se define la localización de un objeto espacial (representado mediante punto, vector, área, volumen) en la tierra de acuerdo a su nombre propio (topónimo) y a otras características descriptivas. Por lo tanto, se puede hablar de georreferenciacion de un texto, un documento, una imagen, etc. La geocodificación de documentos ha sido estudiada intensivamente dentro del contexto de la recuperación de la información geográfica, *Geographic Information Retrieval* (GIR) [35, 36], incluyendo la búsqueda en la Web [37, 38]. En el texto pueden aparecer referencias a múltiples lugares distintos, es decir, topónimos. La investigación en la resolución de topónimos, *Toponym Resolution* (TR)[35], se centra en la georreferenciación de los topónimos en el texto. La eficacia de esta tarea depende, por un lado, de una base de referencia y, por otro lado, del algoritmo usado. El lugar puede tener varios nombres (p. ej., un nombre no oficial) y también puede estar en otros idiomas. Además, los nombres de lugares y sus *footprints* cambian con el tiempo lo que puede resultar en una base de referencia incompleta. El algoritmo debe tener en cuenta la ambigüedad: 1) las palabras comunes deben estar diferenciadas de los nombres propios (ambigüedad *geo/non-geo* [39]), y 2) el mapeo entre los nombres y las localizaciones es ambiguo (p. ej. hay alrededor de 40 "London" en el mundo). Existen diversas aproximaciones, como las taxonomías simples basadas en *gazetteers* [39], las ontologías más complejas de lugares [40] o aplicando otros nombres en el texto para la mejora de la desambiguación de topónimos [41] .

Las herramientas de georreferenciación más comunes son los *geocoders* que suelen ser usados para la georreferenciación de direcciones, nombres de lugares de interés, etc. Actualmente existen diversos servicios de *geocoder* (p. ej., *Google Geocoder* (*Google Geocoding* API[25]), *Yahoo Geocoder* (*Yahoo! PlaceFinde* API[26]), *Geonames*[27]), los cuales se pueden acceder directamente vía API mediante solicitudes HTTP y obtener respuestas en formatos comunes o estándares como XML, JSON o KML. Los *geocoders* Web se diferencian entre ellos en su nivel de detalle, cobertura, accesibilidad o precisión. También se diferencian en su modo de acceso: libre, restringido o de pago. Un *geocoder* recibe como entrada un texto corto y devuelve un listado ordenado de posibles lugares. La eficacia de un *geocoder* depende de los datos de referencia y del algoritmo usado [42]. Por ejemplo,

[22] http://nlp.stanford.edu/ner/index.shtml
[23] http://incubator.apache.org/opennlp/
[24] http://cogcomp.cs.illinois.edu/page/software_view/4
[25] http://code.google.com/intl/es-ES/apis/maps/documentation/geocoding/
[26] http://developer.yahoo.com/geo/placefinder/
[27] http://www.geonames.org/

en el caso de *Google Geocoder*, las heurísticas usadas tienen en cuenta varios elementos: i) las características generales de las entidades geográficas (su importancia, por ejemplo, da la prioridad a las entidades geográficas que corresponden al más alto nivel de la organización territorial; ii) el contexto del usuario (vía IP o KEY), lo cual tiene influencia en los resultados: 1) el idioma del resultado, 2) da la preferencia a los lugares relacionados con la localización del usuario (según los resultados obtenidos, esto suele ser visible principalmente para las entidades de bajo nivel de organización territorial como son las calles).

2.2.3. Transformación de Modelo

La transformación de modelo se puede definir como *la generación automática de uno o múltiples modelos destino desde uno o múltiples modelos origen, de acuerdo a una descripción de la transformación*. [43].

La transformación de modelo puede tener diversas aplicaciones dentro del desarrollo basado en modelos, como por ejemplo: modificación, creación, adaptación, fusión, entrelazado o alteración de modelos. En todas estas tareas es un factor común la reutilización de la información capturada en los modelos. Los usos de la transformación de modelo se pueden agrupar bajo los siguientes aspectos:

- Síntesis: refinamiento que añade detalles al modelo. Movimiento desde un nivel alto de abstracción a un nivel más bajo.
- Integración: incorporación de diferentes herramientas de desarrollo. Fusión de modelos.
- Análisis y simulación: cálculo de métricas, como por ejemplo de similitudes e integración de tecnologías externas de análisis más complejo.
- Optimización: mejora de una o algunas propiedades no funcionales.

En la figura 3 se ilustran los componentes que conforman un proceso de transformación de modelo [44].

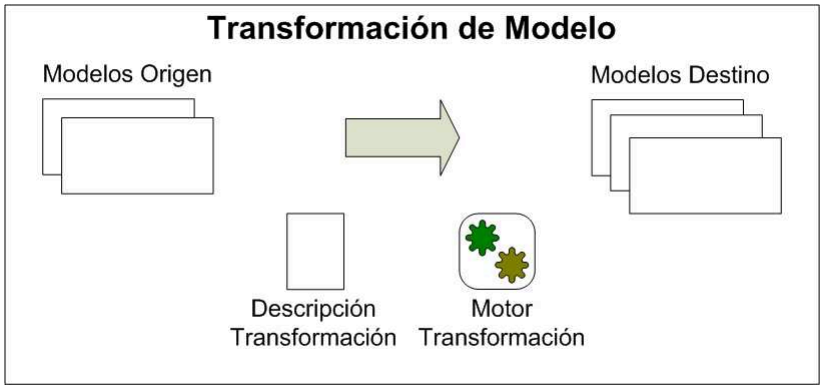

Figura 3: Componentes de la transformación de modelo

- Modelo origen (*source*): un modelo puede tomar el rol de un modelo origen si éste es la entrada de la transformación. Esta entrada puede estar conformada por uno o más modelos origen.
- Descripción de la transformación: expresa como uno o más modelos origen son transformados en uno o más modelos destino y está escrita en un lenguaje de transformación de modelos. Mediante el uso de las reglas de transformación de modelo se describe cómo un fragmento del modelo origen puede transformarse en un fragmento del modelo destino.

- Motor de la transformación: es la herramienta que ejecuta o interpreta la descripción de la transformación sobre el modelo origen para producir el modelo destino. Un motor de transformación ejecuta los siguientes pasos:
  - Identifica los elementos en el modelo origen que necesitan ser transformados.
  - Para cada uno de los elementos identificados, produce los elementos de destino asociados.
  - Genera la información de trazabilidad que enlaza los elementos origen y destino.
- Modelo destino (*target*): un modelo puede tomar el rol de modelo destino si éste es la salida de la transformación. Esta salida puede estar conformada por uno o más modelos destino.

Un problema de transformación de modelo (es decir, un problema que se quiere resolver mediante el uso de una transformación de modelo) se puede clasificar [45, 46] de acuerdo a las siguientes características:

- Cambio de abstracción:
  - Refinamiento vertical: crecimiento del nivel de detalle del modelo origen o decrecimiento del nivel de detalle del modelo origen.
  - Refinamiento horizontal: cambio en la representación del modelo pero no su nivel de abstracción.
- Cambio de metamodelo: un metamodelo de un modelo X describe la estructura que el modelo X debe seguir para que sea válido.
  - Transformación endógena: los metamodelos origen y destino son los mismos. Sólo se cambia una parte específica del origen.
  - Transformación exógena: se efectúa un mapeo de conceptos entre diferentes metamodelos (traducción).
- Soporte al número de modelos:
  - Un modelo: el mismo modelo para el origen y el destino. *(in-place*).
  - Dos modelos: distintos modelos para el origen y el destino. El modelo destino sólo contiene información expresamente generada.
  - Varios modelos origen: se combinan en un solo modelo destino o varios modelos destinos opcionalmente referenciados.
- Soporte al tipo de destino:
  - Modelo a modelo: se realiza un mapeo entre elementos del modelo origen y el modelo destino.
  - Modelo a texto: se realiza un mapeo entre un elemento del modelo origen y un fragmento arbitrario de texto.
- Preservación de propiedades:
  - Semántica: no se cambia el significado del modelo, pero se mejora la estructura y calidad del modelo.
  - Comportamiento: las restricciones explícitas o implícitas del comportamiento en el modelo de origen permanecen completas en el modelo de destino.
  - Sintaxis: transformación horizontal endógena que no cambia la sintaxis abstracta del modelo origen.

Capítulo 3. Diseño y Arquitectura del Sistema

Este capítulo está dedicado a presentar un sistema que permite caracterizar de forma automática, metadatos de recursos Web. La ventaja principal de la arquitectura es su aspecto extensible, factible de incorporar el soporte a diferentes tipos de recursos Web y la flexibilidad del sistema para facilitar la adición de nuevas características.

Una visión general del sistema, funcionalidad, ejemplo de uso y su arquitectura se presentan a continuación en el punto 3.1. Una descripción explícita de los diferentes componentes del sistema de generación de metadatos se describe en el punto 3.2, seguida de la descripción de una implementación realizada de generadores y extractores en el punto 3.3.

3.1. Arquitectura

La funcionalidad general del sistema puede ser descrita de la siguiente manera (figura 4): el sistema recibe como entrada un documento de metadatos del recurso, esto es, un documento con formato XML que describe a un recurso Web. La descripción se basa en un esquema (descripción formal de un modelo) conocido y que contiene un enlace válido a un recurso Web (URL). El sistema detecta el tipo y el formato del recurso relacionado. Luego, algunos metadatos seleccionados pueden ser extraídos desde el recurso (conjunto de metadatos). Varios conjuntos de metadatos pueden ser extraídos de acuerdo a diferentes heurísticas. Estos conjuntos de metadatos serán luego evaluados y usados para generar la descripción del recurso.

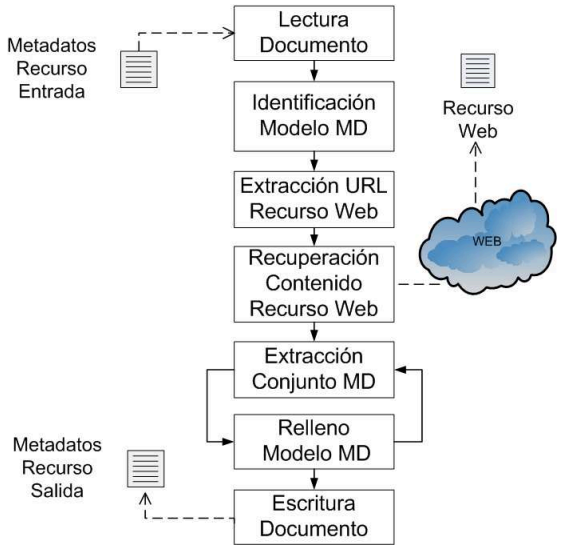

Figura 4: Funcionalidad general del sistema

El modelo de documento de entrada tiene que seguir uno de los esquemas previamente registrado, lo cual permite que el sistema funcione con el documento de entrada. Como el sistema detecta automáticamente el tipo y el formato del recurso Web, éste soporta la mayoría del los tipos MIME, centrándose en los más populares. Para ello se aplican algunas heurísticas adicionales para verificar el tipo de recurso. Esto es especialmente útil en el caso de recursos Web

lógicos (servicios Web) como OWS, por ejemplo, cuando son descritos a través de loe elementos de un catálogo OGC. El sistema puede aplicar diferentes lógicas para analizar el recurso Web y extraer un conjunto de elementos descriptivos. Estos conjuntos de metadatos pueden describir características temáticas (por ejemplo, información geográfica, información acerca del autor o contenido) o depender del tipo de recurso (por ejemplo, el nombre del fichero o su protocolo). La información recopilada se utiliza para satisfacer las carencias en la descripción del documento de entrada, mejorarlo, o validarlo, generando un informe con las diferencias y recomendaciones (comparando los valores desde el documento de entrada con los valores obtenidos). Diferentes algoritmos pueden ser aplicados para realizar estas tareas, por ejemplo, un simple mapeo entre campos correspondientes, o algoritmos más complejos que analicen los conjuntos de metadatos extraídos desde los recursos hijos (otros recursos accesibles desde el propio recurso).

Durante el diseño de la arquitectura del generador de metadatos, el mayor énfasis ha tenido lugar en la extensibilidad y la flexibilidad del sistema para facilitar la adición de nuevas características. Por lo tanto, la idea de extensibilidad del sistema contempla el soporte de una variedad de:

- esquemas de metadatos de documentos de entrada que describen a los recursos Web (por ejemplo, perfiles de especificación DC o ISO de OGC CSW, respuesta de la solicitud de la operación OWS *getCapabilities*),
- tipos de recursos Web que están siendo descritos. Ellos incluyen el documento en que se basa el recurso (texto puro, HTML, KML, etc.) o la lógica de recursos Web (i. e., un geoportal o OWS),
- conjuntos de metadatos que podrían ser extraídos de un recurso Web,
- algoritmos de extracción que se aplicarán sobre un recurso Web y
- algoritmos de generación que se aplicarán sobre conjuntos de metadatos extraídos para producir la descripción de salida del recurso.

En la Figura 5 se presentan los principales componentes de la arquitectura propuesta.

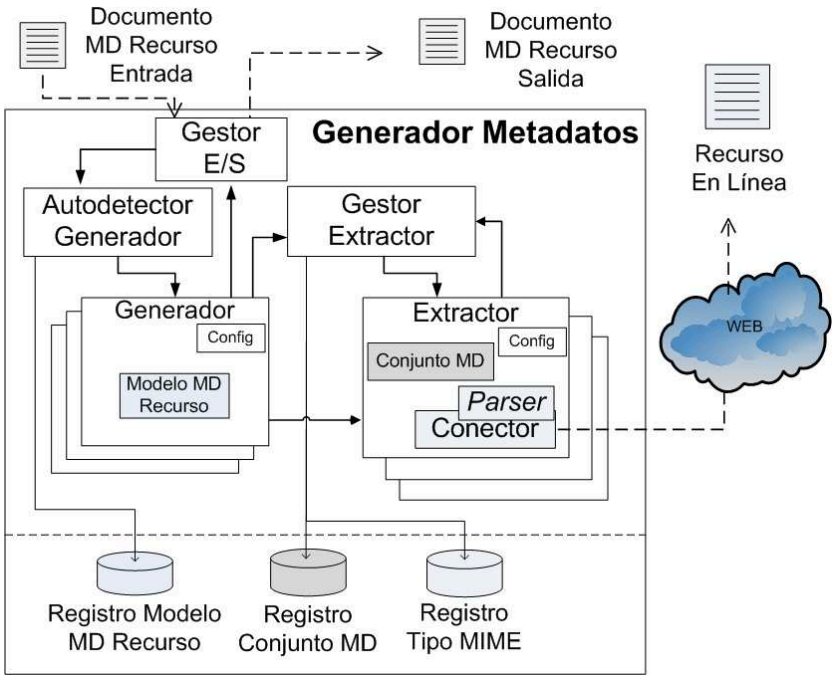

Figura 5: Arquitectura del sistema

3.2. Componentes del Sistema

3.2.1. Gestor Entrada y Salida

El Gestor de Entrada y Salida gestiona la entrada al sistema del documento de metadatos del recurso (Documento MD Recurso Entrada), un documento con formato XML cuyo modelo se ajusta a uno de los esquemas soportados por el sistema y que han sido previamente registrados en un repositorio de esquemas (Registro Modelos MD Recurso). El componente Generador delega a este componente la creación del documento de metadatos de salida del recurso (Documento MD Recurso Salida) de acuerdo al esquema del documento de entrada.

3.2.2. Autodetector del Generador

El Autodetector del Generador analiza el documento de entrada y selecciona el componente de generador apropiado (Generador) de acuerdo al esquema definido en el propio documento. El generador seleccionado conoce como trabajar con este esquema, y opera sobre un modelo interno de descripción del recurso (Modelo MD Recurso) que comprende a dicho esquema.

3.2.3. Generador

El Generador extrae el URL del recurso desde el documento de entrada y delega la extracción de los conjuntos de metadatos al componente Gestor del Extractor. El comportamiento del Generador puede variar dependiendo de los requerimientos del usuario (definidos en un fichero de configuración). Por ejemplo, su lógica puede estar definida mediante varios ficheros de configuración, esto es, en un fichero se puede definir el tipo de conjuntos de metadatos y el orden en el cual ellos deberán ser extraídos, y en un segundo fichero de configuración se puede definir el mapeo de los nombres de metadatos origen con su modelo interno de descripción del recurso.

3.2.4. Gestor del Extractor

Según el conjunto de metadatos solicitado (Conjunto MD) y el tipo de recurso Web, el Gestor del Extractor invoca al componente Extractor apropiado.

3.2.5. Extractor

Un Extractor recupera el contenido del recurso Web (Conector) y retorna el conjunto solicitado de metadatos relacionados con el recurso. Los extractores pueden utilizar diferentes estrategias para la extracción del mismo conjunto de metadatos (Analizador Sintáctico o *Parser*). La estrategia predeterminada está definida mediante un fichero de configuración y puede ser modificada por el usuario. Un extractor puede a su vez estar compuesto de dos o más extractores, lo cuales son gestionados a través del componente Gestor del Extractor.

Capítulo 4. Caso de Uso

4.1. Introducción

Este trabajo se dedicó al desarrollo de una arquitectura para la generación automática de metadatos para recursos Web. Para el estudio de un caso de uso, se desarrolló una aplicación prototipo dedicada a la extracción de las propiedades usadas comúnmente por los servicios de catálogo OGC (OGC CSW [47]) para describir un recurso, el registro CSW (Tabla 4). Como entrada, al sistema se le ha proporcionado un documento del esquema del registro CSW, con el elemento *source* que contenía el URL del recurso Web que debe estar descrito.

Elemento CSW	Elemento META de HTML	Definición	O/I
contributor	`DC.contributor`	Entidad responsable contribuciones	O

coverage	`DC.coverage.*`
	`geographic-coverage`
	`ICBM`
	`geo.position,`
	`geo.placename`
	`geo.region`
		Extensión espacial o alcance del contenido del recurso (traducido a *Bounding Box*)	O

creator	`DC.creator` `author, webauthor`	Entidad responsable de creación del contenido del recurso	O
date		Fecha de creación o actualización **Valor por defecto**: fecha y hora del sistema (ISO 8601)	I
description	`DC.description` `description`	Descripción breve del contenido del recurso	O

format	`DC.format`
	`content-type`
	`(http-equiv)`
		Manifestación física o digital del recurso	O

identifier		**Valor por defecto**: Referencia única del registro (GUID)	I
language	`DC.language` `language`	Lenguaje del contenido del registro	O
publisher	`DC.publisher` `publisher`	Entidad responsable de la publicación del recurso	O
relation		Nombre de relación entre recurso y el recurso relacionado referenciado mediante el elemento "*source*". (**Valor por defecto:** "URL")	I
rights	`DC.rights` `rights`	Licencia de los derechos o licencia de uso del recurso	O
source		Recurso relacionado de donde se deriva descripción (**Valor por defecto**: URL del Recurso Web)	I
subject	`DC.subject` `keyword, keywords`	Temática del contenido del recurso	O
title	`DC.title` `application-name`	Nombre del recurso	I
type		Naturaleza o género del recurso. **Valor por defecto**: "Geoportal"	I

Tabla 4: El mapeo entre elementos comunes del registro CSW y elementos META de HTML (O/I – Opcional/Imperativo)

4.2. Estudio de soluciones

Para el desarrollo del prototipo se revisaron y estudiaron algunas propuestas de extractores, factibles de integrar dentro la arquitectura del sistema propuesto, para la recolección y/o generación de metadatos relacionados con cada una de las propiedades DC asociadas a un recurso CSW. En la tabla 5 se relaciona cada propiedad de metadato con uno o más extractores (indicados con letras) y algunas reglas simples, expresadas en el lenguaje de la Teoría de Conjuntos, para la combinación entre los modelos de los conjuntos de metadatos de cada extractor.

Metadato	A	B	C	D	E	F	G	H	I	J	K	L	M	Reglas
conributor	✓													
coverage	✓						✓	✓	✓			✓	✓	$\neg A \rightarrow M$ $\neg M \rightarrow I$ $\neg I \rightarrow H$ $\neg H \rightarrow L$
creator	✓													
date	✓													Valor por defecto: Fecha de creación (ISO 8601)
description	✓					✓								$\neg A \rightarrow F$
format	✓													
identifier	✓													Valor por defecto: GUID
language	✓	✓												$\neg A \rightarrow B$ (En *Tika*: $B \subset A$)
publisher	✓						✓							$\neg A \rightarrow \{m \in G: m \approx$ "Organización"$\}$
relation	✓													Valor por defecto: "*URL*"
rights	✓									✓				$\neg A \rightarrow J$
source	✓													Valor por defecto: URL del recurso Web
subject	✓				✓						✓			$\neg K \rightarrow F$ $\neg E \rightarrow A$
title	✓		✓	✓										$\neg A \rightarrow C$ $\neg C \rightarrow D$
type	✓													Valor por defecto: "*Geoportal*"

Tabla 5: Propuestas de extractores y reglas de combinación (✓: Posible fuente de información o mapeo)

A continuación se explican brevemente cada uno de los extractores propuestos:

A. <META>: Extracción del contenido de los atributos `name`, `http-equiv` y `content` de los elementos `META` de un documento HTML. Adicionalmente, se extrae el contenido del elemento `TITLE`. [33]

B. *N-Gram*: Categorización de un texto basado en un algoritmo de cálculo de frecuencias de palabras aplicado también a caracteres individuales. Se aplica para la detección del lenguaje del texto. [33,48]

C. <H1>: Extracción del contenido encerrado entre los elementos `H1` (encabezado) del cuerpo de un documento HTML. [49]

D. *Text*50: Extracción de una secuencias de palabras para los 50 primeros caracteres del contenido del cuerpo de un documento HTML. [49]

E. *PhraseRate*: Extracción de palabras claves (2-5 palabras) a partir de documentos HTML bien escritos y orientados a temas específicos. [50]

F. *AutoAnnotator*: Extracción del párrafo simple que mejor representa el resumen del cuerpo de un documento HTML. [51]

G. NER: Extracción de nombres de entidades (Localidad) contenidos en un documento HTML o en texto plano, basado en un algoritmo NLP. [52]

H. <IMG>NER: Extracción de nombres de entidades (Localidad) a partir del contenido (texto alternativo) del atributo `alt` del elemento `IMG` (imagen) de un documento HTML.

I. <A>NER: Extracción de nombres de entidades (Localidad) a partir del contenido (texto visible) que encierra el elemento `A` (anclaje o enlace) de un documento HTML.

J. *License*: Extracción del URL a partir del contenido del atributo `href` de los elementos `A` (anclaje) y `LINK` cuyo contenido del atributo `rel` sea igual a "*Copyright*" o "*Licence*" de un documento HTML. [53]

K. LCSH: Clasificación LCSH (*Library Congress Subject Heading*) basado en algoritmo de *Naive Bayes* localmente ponderado del texto de un documento HTML. [54]

L. <BODY>NER: Extracción de nombres de entidades (Localidad) a partir del contenido (texto visible sin enlaces ni imágenes) del cuerpo de un documento HTML.

M. <META>NER: Extracción de nombres de entidades (Localidad) a partir del contenido de los elementos `META` de un documento HTML.

4.3. Desarrollo de un prototipo

El prototipo fue desarrollado con el objetivo de generar los metadatos de un registro CSW para recursos Web. Por esta razón, el primer paso es la recolección de los metadatos existentes en la cabecera (`HEAD`) del propio documento HTML (elementos `META`) y del título del mismo (elemento `TITLE`). Es necesario tener en cuenta que en la mayoría de los recursos Web, los nombres utilizados para un mismo metadato suelen ser diversos y, generalmente, no suelen estar ajustados a ninguna normativa o estandarización. Por esta razón se hace necesario realizar un mapeo (traducción) entre los nombres de metadatos frecuentemente usados y las propiedades (*Dublin Core*) que conforman a un registro CSW. Este mapeo debe de ser dinámico, por lo que debe ser gestionado desde un fichero de configuración externa, fácilmente editable y con opciones para asignar valores de peso, según la prioridad requerida, para una misma propiedad DC.

En el caso de la ausencia de algún metadato específico en la cabecera (`HEAD`) del documento HTML, es necesario analizar el contenido del cuerpo (`BODY`) para intentar extraer o inferir información asociada a dicho metadato y definir algunas reglas básicas de transformación de modelo para decidir bajo que condiciones se debería realizar este análisis.

Para efectos de la implementación del prototipo se han elegido las siguientes propuestas de extractores (explicadas en la sección anterior): (A) <META>, (G) NER, (H) <IMG>NER, (I) <A>NER, (J) *License,* (L) <*BODY*>*NER* y <META>NER.

4.4. Implementación

El prototipo para el caso de uso ha sido implementado en el lenguaje *Java* con un aprovechamiento de parte del trabajo del proyecto *Apache Tika*. En el Anexo IV se describe el

conjunto de interfaces genéricos que soportan el diseño del prototipo desarrollado para la implementación de la arquitectura propuesta.

Para la implementación del conjunto de metadatos se ha extendido la clase original de *Apache Tika* a través de la clase *CompoundMetadataImpl* , donde se ha mejorado la estructura de datos y su gestión para contemplar más información relacionada con cada propiedad registrada, además de su propio valor: la propiedad padre (para los casos de propiedades compuestas), el elemento del documento HTML (para discriminar su ubicación dentro del documento: `TITLE`, META y `BODY`), un valor alternativo y la frecuencia de aparición de un mismo valor (repetición).

Al modelo del conjunto de metadatos extraídos por un extractor se le aplica una transformación exógena (traducción) hacia el modelo interno mediante la realización de un mapeo entre los nombres de metadatos de origen y las propiedades del modelo interno (clase *ArrangerCSW*), de acuerdo con la tabla 4 (sección anterior). En el caso del generador con múltiples extractores, la regla básica establecida para la invocación del extractor específico de una propiedad en particular es la ausencia de valor para dicha propiedad en el modelo interno del recurso.

La tabla 6 resume la implementación de los extractores según la funcionalidad de los siete extractores señalados en la sección anterior. La clase *AdaptHtmlExtractorImpl* ha sido implementada para la extracción de metadatos tanto desde la cabecera (`HEAD`) como desde el cuerpo (`BODY`) de un documento HTML, y la clase *NERExtractorImpl,* para la extracción de nombres de entidades (geográficas) de todo el documento HTML (o desde un texto plano) mediante el uso de una herramienta NER (*Stanford* NER). La primera de estas dos clases permite adaptarse a diferentes clases de controladores (*handlers*) para la extracción de propiedades según sea el tipo de conjunto de metadatos especificado. Mediante el desarrollo de los extractores compuestos, que combinan las dos clases anteriores, se ha implementado la funcionalidad de extracción de nombres geográficos (aplicando una herramienta NER) desde el texto visible de todos los enlaces extraídos (*AnchorNERExtractorImpl*), desde el texto alternativo de todas las imágenes extraídas (*ImgNERExtractorImpl*), y desde el texto visible sin enlaces ni imágenes (*BodyNERExtractorImpl*) en el cuerpo (`BODY`) de un documento HTML, desde el contenido de todos los metadatos geográficos (*GeoMetaNERExtractorImpl*) y desde el contenido de todos los metadatos (no geográficos) (*MetaNERExtractorImpl*) en la cabecera (`HEADER`) del mismo documento.

Extractor	Nombre de Clase	*Handler*	S/C	Observaciones
A	*AdaptHtmlExtractorImpl*	*MetaHtmlHandlerImpl*	S	
G	*NERExtractorImpl*	-	S	
J	*AdaptHtmlExtractorImpl*	*LicenseHtmlHandlerImpl*	S	
L	*BodyNERExtractorImpl*	-	S	
M	*GeoMetaNERExtractorImpl*	*MetaHtmlHandlerImpl*	C	Sobre el resultado de extractor A (inicializado con *handler* indicado) se aplica el extractor G
N	*MetaNERExtractorImpl*	*MetaHtmlHandlerImpl*	C	Sobre el resultado de extractor A (inicializado con *handler* indicado) se aplica el extractor G
H	*ImgNERExtractorImpl*	*ImgHtmlHandlerImpl*	C	Sobre el resultado de extractor A (inicializado con *handler* indicado) se aplica el extractor G
I	*AnchorNERExtractorImpl*	*AnchorHtmlHandlerImpl*	C	Sobre el resultado de extractor A (inicializado con *handler* indicado) se aplica el extractor G

Tabla 6: Detalles de la implementación de los extractores seleccionados para el prototipo (S: Simple, C: Compuesto).

Para la gestión de los eventos que controlan la extracción de los nombres de propiedades y sus valores desde el documento HTML, y la asignación de éstos dentro del conjunto de metadatos se han implementado las siguientes clases de controladores (*handlers*): *MetaHandlerImpl*, para la extracción de los elementos `META` (nombre y valor) y el elemento `TITLE` (valor) desde la cabecera (`HEAD`); *AnchorHtmlHandlerImpl*, para la extracción de los elementos `A` (URL y texto visible) desde el cuerpo (`BODY`); *ImgHtmlHandlerImpl* , para la extracción de los elementos `IMG` (URL y texto alternativo) desde el cuerpo; y *LicenseHtmlHandlerImpl* para la extracción del elemento `LINK` (URL) desde la cabecera (`HEAD`), y del elemento `A` - *Anchor* - (URL) desde el cuerpo (`BODY`) cuando el valor del atributo `rel` para ambos elementos sea "*Copyright*" o "*License*" respectivamente.

Con respecto a los generadores, se han implementado dos tipos de generadores: un generador simple (clase *MetadataGeneratorProviderCSWImpl*) para la extracción de nombres y valores de metadatos desde la cabecera (`HEAD`) de un documento HTML y otro generador múltiple (clase *MetadataMultipleGeneratorProviderCSWImpl*) para la invocación conjunta de varias clases de extractores que, además de actuar sobre la cabecera (`HEAD`), actúan sobre el cuerpo (`BODY`) del documento HTML. En ambos casos, se trabaja con un modelo interno de descripción de las propiedades del recurso (clase *CSW*), que comprende el esquema de un registro CSW, cuya gestión de lectura de propiedades del documento XML de entrada y de creación del documento XML de salida se ha implementado mediante sus respectivas clases (*ReaderCSW* y *WriterCSW)*.

4.5. Experimentos y resultados

En esta sección se describen los experimentos realizados para validar el prototipo del caso de uso y se hace una discusión de los resultados obtenidos. En primer lugar se describe el corpus de datos que se usó en el experimento.

4.5.1. Corpus de datos

El corpus usado para los experimentos está compuesto por enlaces (URL) de geoportales. Al principio del trabajo, el corpus se había recopilado desde la Web mediante el uso de los buscadores Web existentes para localizar geoportales o proyectos científicos con temática relacionada con la información geográfica (hidrografía, geología, etc.), extrayendo también los enlaces de los geoportales encontrados o las referencias encontradas en los proyectos. No obstante, las propias características dinámicas de la Web han tenido mucha influencia en la definición del corpus final usado en el trabajo. En primer lugar, muchos de los enlaces (URL) de las páginas Web de los geoportales que se tuvieron en cuenta al comienzo del proyecto no se pudieron analizar por estar obsoletas o inaccesibles.

Por esta razón, para validar el prototipo del caso de uso se llevó a cabo un par de experimentos sobre una muestra representativa de páginas Web (ver Anexo III) vinculadas con diversas iniciativas, de ámbito global, regional, nacional y local, relacionadas con las infraestructuras de datos espaciales, recopiladas por la *Global Spatial Data Infrastructure Association* (GSDI) en su sitio Web[28]. El objetivo principal de esta asociación es el de promover la cooperación y la colaboración internacional en apoyo de las infraestructuras de datos espaciales locales, nacionales e internacionales, que permitan a las naciones abordar mejor las cuestiones sociales, económicas y ambientales que tengan una importancia apremiante. Una vez definido en un momento dado el corpus para los experimentos resultó que durante el tiempo en que se desarrollaba el prototipo, algunos enlaces (URL) de páginas Web de geoportales quedaron inaccesibles. Además, aunque el

[28] http://www.gsdi.org/

prototipo está preparado para procesar las páginas Web creadas con errores de sintaxis, se encontraron algunos ejemplos de páginas Web que fueron imposibles procesar debido a su mal diseño. Estos enlaces (URL) de páginas Web fueron eliminadas del corpus. Estas decisiones se reflejan en el Anexo V.

4.5.2. Primer experimento

Durante el primer experimento, el prototipo fue aplicado a la extracción automática de los elementos `META` de cada una de las páginas Web de la muestra seleccionada. La Tabla 4 presenta el mapeo conceptual aplicado en este experimento, entre los elementos de metadatos generados (registro CSW) y los elementos `META` del documento HTML. Además, el elemento `TITLE` se añadió como parte de este mapeo y se ha utilizado cuando el contenido del elemento `META` asociado al título de la página del documento no se puede extraer. En la tabla 7 se muestran los resultados obtenidos en este primer experimento.

Metadato	% Encontrado[29]	Media de Valores[30]
contributor	0%	0
coverage	2%	2
creator	26%	1,1154
description	35%	1,0857
language	41%	1,0732
publisher	9%	1
rights	11%	1
subject	41%	1,0732
title	93%	1,1290

Tabla 7: Resultados del experimento 1 (Total de elementos del corpus: 122; no procesados[31]: 18%)

El generador utilizado en este experimento se ha caracterizado por generar un solo valor por metadato para los casos donde su extractor consigue extraer valores de los elementos `META` (y `TITLE`) de las páginas Web procesadas. Además, estos valores no suelen estar duplicados.

Este generador funciona muy bien para obtener *title* en casi la totalidad de los casos. Además, en casi de la mitad de los casos, los elementos `META` proporcionan información sobre *subject* y *language*, un tercio sobre *description*, una cuarta parte sobre *creator* y apenas una décima parte sobre *publisher* y *rights*. No se genera ningún valor para *contributor*.

Con respecto a *coverage*, propiedad importante para describir la extensión geográfica de los recursos, ha sido identificada en apenas el 2% de las páginas Web procesadas (usando metadatos geográficos de posición). La dedicación a la información geográfica de los recursos Web analizados permite asumir que *coverage* se podría estimar de forma automática desde el contenido de estos recursos. Por esta razón, el segundo experimento se enfocó principalmente a este aspecto.

[29] Metadatos encontrados en las páginas Web procesadas exitosamente.
[30] Media del número de valores recogidos por elemento para describir el metadato
[31] No procesadas a causa de servidor no accesible o problemas en la carga de la página Web.

4.5.3. Segundo experimento

El segundo experimento ha sido dedicado a la mejora de la generación de metadatos producida en el experimento 1. El enfoque principal ha sido la parte geográfica. Por lo tanto ha sido necesaria la identificación y la definición de las heurísticas para reconocer y tratar correctamente los nombres geográficos encontrados en los diferentes elementos de la página Web para la estimación de la extensión geográfica (*Bounding Box*) de la página. Como primer paso, se ha realizado un análisis de forma manual de las características de un conjunto de ejemplos de páginas principales de geoportales. Una muestra representativa ha sido seleccionada del corpus (3 globales, 2 regionales, 4 nacionales y 5 locales). En el Anexo V.I. se presenta el detalle del análisis realizado. Se han analizado los siguientes elementos de la página Web: 1) Posición (longitud/latitud) y nombres geográficos de los elementos `META` geográficos, 2) nombres geográficos de los elementos `META` no geográficos y el elemento `TITLE`, 3) nombres geográficos del texto visible de los enlaces (`A`) del elemento `BODY`, 4) nombres geográficos del texto no visible (`alt`) de las imágenes (`IMG`) del elemento `BODY`, 5) nombres geográficos del texto visible del elemento `BODY` sin 3) y sin 4).

A base del análisis se puede estimar una heurística general aplicable independientemente del tipo del geoportal.

> *(H1) Heurística General*: Si se encuentran metadatos geográficos (1), sus valores son prioritarios dando la preferencia a los metadatos de posición (latitud/longitud). En caso de su ausencia, se utilizan los nombres geográficos encontrados (con mayor frecuencia) en los elementos `META` no geográficos (2) para generar la extensión geográfica *(Bounding Box*).

La heurística H1 es aplicable en el caso de la existencia de elementos `META` y si no hay metadatos geográficos o no han sido identificados nombres geográficos dentro del texto de los elementos `META`, se debe acudir al contenido de la página Web. El proceso de estimación de la extensión geográfica que se debe aplicar a los conjuntos de nombres geográficos encontrados en el contenido de la página Web (elementos (3), (4) y (5)) podría variar dependiendo de la clasificación (ámbito) del geoportal. En el caso de los *geoportales globales* se ha observado que la fuente principal de los nombres geográficos es el (5). Además, hay mucha variedad de nombres de países que pertenecen a las regiones de diferentes partes del mundo. En el caso de los *geoportales regionales* se ha observado que (3) y (5) son la fuente principal de los nombres geográficos y los nombres de países pertenecen a una región geográfica. En el caso de los *geoportales nacionales*, la principal característica es la escasez de los nombres geográficos identificados en (3), (4) y (5). Suele aparecer sólo un nombre de país, eventualmente el nombre de su capital. El análisis también identificó en este caso que el código de país que aparece en el dominio del *source* (URL) suele ser también un buen indicador del país. En el caso de los *geoportales locales* se ha observado que (3), (4) y (5) son la fuente de los nombres geográficos que deben ser tratados. El nombre geográfico de mayor frecuencia dentro del conjunto compuesto por (3), (4) y (5) representa una región de un país (administrativo y/o geográfico) el cual describe la extensión geográfica del geoportal. Los nombres de países son escasos, y si aparecen, uno de ellos suele contener el identificador de región.

Hay que tener en cuenta que no se conoce el tipo de geoportal a priori. Por lo tanto, la heurística aplicada al contenido (H2) debe intentar ser independiente de ello. Se ha propuesto una heurística muy simple:

> *(H2) Simple:* La extensión geográfica de la página Web se estima a base del nombre geográfico de mayor frecuencia. En caso de que no exista un único nombre geográfico se

agrupan los nombres geográficos según la organización territorial a la cual pertenecen empezando desde el menor nivel hacia arriba.

Los detalles del proceso de análisis y los algoritmos de la dos heurísticas está detallado en Anexo V.II.

El prototipo fue ampliado con la finalidad de mejorar la generación de los metadatos con un nuevo generador desarrollado para hacer uso de varios extractores adicionales. Para la extracción de nombres geográficos han sido desarrollados otros cinco extractores que son capaces de identificar los nombres geográficos dentro las diferentes partes del documento HTML, mediante la aplicación de un algoritmo de procesamiento de lenguaje natural (NLP):

1) NER sobre el texto del contenido de los elementos `META` (geográficos)
2) NER sobre el texto del contenido de los elementos `META` (no geográficos)
3) NER sobre el texto visible de los enlaces del elemento `BODY`
4) NER sobre el texto no visible de las imágenes del elemento `BODY`
5) NER sobre el texto visible del elemento BODY (sin el texto visible de los enlaces y sin el texto no visible de las imágenes).

El generador ha sido ampliado adicionalmente con un componente para la estimación de la extensión geográfica, que se basa en los resultados de estos extractores mencionados y devuelve un *Bounding Box* (área delimitada por dos longitudes y dos latitudes). Por razones de evaluación del experimento, también devuelve las entidades geográficas que sirvieron para la generación de este *Bounding Box*. En el Anexo V.II se encuentran más detalles de la implementación.

Se realizaron dos pruebas. En la primera de ella fue aplicada sólo la heurística H1 que trabaja con la parte `META` del documento HTML. En una segunda ejecución de la prueba, las dos heurísticas, H1 (*General*) y la H2 (*Simple*), han sido usadas. Los resultados han sido evaluados manualmente. El resultado "Correcto" significa que la entidad geográfica asignada por el componente de generación del *Bounding Box* es igual a la extensión geográfica estimada de manera manual. "Aceptable" significa que la entidad geográfica asignada es una entidad "padre" (directo) de la entidad asignada manualmente (por ejemplo, "Europa" en lugar de "Slovenia"). El caso inverso se estimó como "Error" (por ejemplo, "Uganda" en lugar de "World"). También había casos donde no se ha producido ningún nombre geográfico "No aplicable (no NG)" en cuyo caso, el componente asigna el *Bounding Box* del "Mundo" (W*orld*). En la tabla 8 se muestran los resultados obtenidos en este experimento.

Resultado	**Heurística**	**%**
Correcto	H1	46,14%
Aceptable (menor precisión)	H1	7,70%
Error	H1	15,39%
No aplicable (no NG)	H1	30,77%
Correcto	H1+H2	53,84%
Aceptable (menor precisión)	H1+H2	15,39%
Error	H1+H2	23,07%
No aplicable (no NG)	H1+H2	7,70%

Tabla 8: Resultados del experimento 2 (Total de elementos del corpus: 122; no procesados: 18%)

Como resultado, se ha observado, que en el caso de aplicar de forma conjunta las heurísticas H1 (*General*) y H2 (*Simple*), se destaca una mejora en los resultados "Correctos" y "Aceptables" de un 7,70% y 7,69% respectivamente.

4.5.4. Discusión

En base a los resultados de los experimentos 1 y 2 se puede observar una mejoría importante en la obtención de la propiedad *coverage*. En el experimento 1 se pudo asignar sólo el 2% de los *Bounding Box* y en el experimento 2, un 92,3% de los *Bounding Box* han sido estimados desde el documento HTML. La validación de los resultados ha indicado un error de un 23,08%. Para los casos donde no se ha podido obtener nombres geográficos (7,7%), se ha asignado el *Bounding Box* del "Mundo" (*World*). Esto puede indicar que usando una heurística muy simple (basada en la frecuencia de nombres geográficos y la agrupación según la pertenencia a una unidad de organización territorial) permite estimar, con un nivel satisfactorio, el *Bounding Box* en casi un 70%.

Hay que destacar, que la heurística final (H1+H2) no ha tenido en cuenta las características deducidas del análisis realizado en los distintos elementos del documento HTML de un geoportal y que sólo se han se han considerado de forma parcial (elemento `META`) en la heurística H1. Aunque se propusieron varias heurísticas que se basaban en las características analizadas del elemento `BODY`, sus resultados no fueron satisfactorios, por lo que se optó por usar una heurística más simple como es la H2. Por lo tanto, el hecho de que al final no se hayan aprovechado las características de los geoportales para el desarrollo de la heurística final permite sospechar que esta podría ser igual de eficaz en otro tipo de documentos HTML y no sólo en las páginas principales de los geoportales.

A los resultados de los experimentos han influido una serie de limitaciones. Por un lado, el prototipo no soporta al contenido dinámico de las páginas Web (*JavaScript*, etc.) y los extractores no obtienen sus datos. Por otro lado, el prototipo desarrollado tiene unas características que provienen de las decisiones tomadas durante su diseño. En primer lugar, estas características dependen de las herramientas externas aplicadas, por ejemplo, la extracción de los nombres geográficos está delegada a la herramienta externa NER seleccionada y al clasificador aplicado (que tiene importancia especialmente en caso de trabajar con páginas plurilingües). Por lo tanto, la precisión de extracción de los nombres geográficos está heredada. En segundo lugar, para la extracción del elemento de *Bounding Box* se ha desarrollado y utilizado una herramienta simple basada en un *geocoder* externo. Esta herramienta utiliza el *geocoder* como fuente de información sobre la organización territorial y para la extracción de *Bounding Box*. También permite solventar la problemática de ambigüedad de topónimos, dar soporte plurilingüe y tratar algunos errores de la herramienta NER. Por lo tanto, su funcionamiento y los resultados dependen fuertemente del *geocoder* aplicado.

Adicionalmente, hay que subrayar que la problemática de la naturaleza dinámica de la Web ha tenido influencia en los experimentos y el trabajo realizado. Por un lado, la volatilidad de los enlaces de las páginas Web (descrita anteriormente en la definición del corpus), y por otro lado, las características dinámicas de la Web también estaban visibles en el hecho de que algunas de las páginas Web analizadas estaban inaccesibles en un momento dado, lo que requería repetir posteriormente los experimentos.

Capítulo 5. Conclusión y Trabajo Futuro

Este trabajo se dedicó al desarrollo de una arquitectura para la generación automática de metadatos geográficos para recursos de Web, con aspecto extensible y flexibilidad para la adición de nuevas características. Para el estudio de un caso de uso, el prototipo desarrollado se ha empleado para la generación de registros OGC CSW que describan a los recursos Web. El primer experimento, realizado para la validación del prototipo, sobre una muestra representativa de páginas Web principales de geoportales, ha demostrado que el principal problema era la extracción de de información relacionada con la extensión geográfica, ya que las páginas Web no suelen contener metadatos geográficos específicos. Por esta razón, para el segundo experimento el prototipo se ha complementado con una herramienta NER que aplica algoritmos NLP para la extracción de nombres de lugares del texto. También ha sido desarrollada una herramienta (integrada como componente) para la estimación de la extensión geográfica (*Bounding Box*) que contemplan los nombres geográficos encontrados dentro de los diferentes elementos de la página Web, sus frecuencias y las relaciones entre ellos. Esta herramienta usa como base un *geocoder* gracias al cual se permite solventar los problemas de la ambigüedad de los topónimos y el soporte al plurilingüismo.

Los resultados del segundo experimento pueden indicar que usando una heurística muy simple (basada en la frecuencia de nombres geográficos y la agrupación según la pertenencia a una unidad de organización territorial) se puede estimar la extensión geográfica, con un nivel satisfactorio, en casi un 70%.

Aunque este prototipo se ha aplicado al dominio de la Web geoespacial, su arquitectura extensible es aplicable a otros contextos que requieran la extracción y generación de metadatos desde otros recursos Web.

Durante el período de la realización de este Trabajo de Fin de Máster se han realizado labores de investigación, desarrollo e innovación que han contribuido a la participación en dos publicaciones de investigación: un artículo [55] en la 9ª Semana Geomática Internacional 2011 (*9th International Geomatics Week*); y un segundo artículo [56] presentado en la Conferencia INPIRE 2011 (*INSPIREd by 2020 - Contributing to smart, sustainable and inclusive growth*).

El trabajo futuro se centrará en otros recursos de la Web geoespacial, como por ejemplo, los documentos de descripción de servicios OGC (respuesta de la solicitud de la operación OWS *getCapabilities*), los ficheros *Shapefiles*, KML y GeoRSS. Se extenderá la implementación actual para poder generar los documentos de metadatos de estos recursos en base al análisis de los recursos vinculados a ellos. Se investigarán los enfoques para mejorar la generación de los metadatos de las páginas Web que utilizan secuencias de instrucciones embebidas (p. ej., *javascripts*) y que generan contenido que debe ser accesible para los extractores. También, utilizando como base los resultados de este trabajo, será necesario investigar y desarrollar las heurísticas de estimación de la extensión geográfica (*Bounding Box*) para otros tipos de recursos Web que se consideren en el futuro. El aspecto plurilingüe será tratado principalmente para usar diferentes clasificadores dependiendo del lenguaje de la página Web, lo que debería de traducirse en una mejora de funcionamiento de la herramienta NER usada.

Capítulo 6. Bibliografía

[1]. Dawes, S.S., Helbig, N., 2010. Information Strategies for Open Government: Challenges and Prospects for Deriving Public Value from Government Transparency, InProc. Electronic Government, 9th IFIP WG 8.5 International Conference, EGOV 2010, Lausanne.

[2]. Franklin, C., Hane, P., 1992. "An introduction to GIS: linking maps to databases," Database. Vol. 15, No. 2 April.

[3]. ESRI, 1998. ESRI Shapefile Technical Description. An ESRI White Paper, http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf (Accedido por última vez en Julio de 2011).

[4]. Wilson, T., 2008. OGC KML Version 2.2.0 (OGC 07-147r2), Open Geospatial Consortium, Inc.

[5]. Portele, C., 2007. OpenGIS Geography Markup Language (GML) Encoding Standard (OGC 07-036), Open Geospatial Consortium, Inc.

[6]. Powell, A., Nilsson, M., Naeve, A., Johnston, P., 2005. Dublin Core Metadata Initiative - Abstract Model (White Paper), DCMI. http://www.dublincore.org/documents/abstract-model/ (Accedido por última vez en Julio de 2011).

[7]. ISO/TC 211, 2003. ISO 19115:2003 Geographic information – Metadata, International Organization for Standardization, Geneva.

[8]. ISO/TC 211, 2007. ISO/TS 19139:2007 Geographic information – Metadata – XML schema implementation, International Organization for Standardization, Geneva.

[9]. Whiteside, A., Greenwood, J., 2010. OGC Web Services Common Standard Version 2.0.0, Open Geospatial Consortium, Inc.

[10]. Ritter, N., Ruth, M., 2000. GeoTIFF Format Specification Revision 1.0, GeoTIFF Working Group. http://www.remotesensing.org/geotiff/spec/geotiffhome.html (Accedido por última vez en Julio de 2011).

[11]. Nebert, D.D., 2004. Developing Spatial Data Infrastructures: The SDI Cookbook. GSDI.

[12]. Turner, A., 2006. Introduction to Neogeography. O'Reilly Media.

[13]. Egenhofer, M.J., Mark, D.M., 1995. Naive Geography. COSIT'95.

[14]. Goodchild, M.F., 2007. Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0, International Journal of Spatial Data Infrastructures Research, Vol. 2.

[15]. Craglia, M., Goodchild, M.F., et al., 2008. Next-Generation Digital Earth: A position paper from the Vespucci Initiative for the Advancement of Geographic Information Science. International Journal of Spatial Data Infrastructures Research.

[16]. Keßler, C., Janowicz, K., Bishr, M., 2009. An agenda for the next generation gazetteer: Geographic information contribution and retrieval. In: International Conference on Advances in Geographic Information Systems 2009 (ACM SIGSPATIAL GIS 2009).

[17]. Page, L. y Brin, S., 1998. The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN Systems, Vol. 30, No. 1-7.

[18]. Wukovitz, L.D., 2001. Using internet search engines and library catalogs to locate toxicology information, Toxicology, Vol. 157, No. 1–2.

[19]. Béjar, R., Nogueras-Iso, J., Latre, M.A. Muro-Medrano P. R., Zarazaga-Soria, F. J., 2009. Digital Libraries as a Foundation of Spatial Data Infrastructures. Handbook of Research on Digital Libraries: Design, Development, and Impact, IGI Global, Singapore.

[20]. Greenberg, J., 2003. Metadata Generation: Processes, People and Tools. Bulletin of the American Society for Information Science and Technology, vol. 29, no. 2, December/January. http://www.asis.org/Bulletin/Dec-02/greenberg.html (Accedido por última vez en Julio de 2011).

[21]. Kalantari, M., Olfat, H., Rajabifard, A., 2010. Automatic Spatial Metadata Enrichment: Reducing Metadata Creation Burden through Spatial. GSDI 12 pre-conference refereed book.

[22]. Ossenbruggen, J., Nack, F., Hardman, L., 2004. That Obscure Object of Desire: Multimedia Metadata on the Web, Part I, IEEE Multimedia, Vol. 11, No. 4, pp. 38-48.

[23]. Nack, F., Ossenbruggen, J., Hardman, L., 2003. That Obscure Object of Desire: Multimedia Metadata on the Web, Part II, IEEE Multimedia, Vol. 12, No. 1, pp. 54-63.

[24]. Foulonneau, M., Riley, J., 2008. Metadata for digital resources: implementation, systems design and interoperability, Chandos Information Professional Series, Oxford.

[25]. Greenberg, J., Pattuelli. M.C., Parsia. B., Davenport Robertson, W., 2001. Author-generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization. Journal of Digital Information.

[26]. Sean A. Golliher, S.A, 2008. Search Engine Ranking Variables and Algorithms, SEMJ.ORG Vol. 1, Supplemental Issue, August. http://www.semj.org/dmdocuments/search_engine_ranking/_algorithms.pdf (Accedido por última vez en Julio de 2011).

[27]. Hickson, I., 2011. HTML5 A vocabulary and associated APIs for HTML and XHTML. Editor's Draft 19 February 2011. W3C. http://dev.w3.org/html5/spec/Overview.html#meta (Accedido por última vez en Julio de 2011).

[28]. Hickson, I., 2011. HTML Living Standard — Last Updated 19 January 2011, WHATWG Web Applications 1.0 specification.

[29]. Polfreman, M., Rajbhandari, S., 2008. MetaTools - Investigating Metadata Generation Tools Final Report, London. http://www.jisc.ac.uk/media/documents/programmes/reppres/metatoolsfinalreport.pdf (Accedido por última vez en Julio de 2011).

[30]. Humphreys J.B.K., 2002. PhraseRate: An HTML Keyphrase Extractor. Technical report, University of California, Riverside. http://infomine.ucr.edu/projects/publications/Humphreys-2002-PhraseRate.pdf (Accedido por última vez en Febrero de 2011.

[31]. Greenberg, J., 2004. Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications. Journal of Internet Cataloging. Taylor & Francis.

[32]. Noufal, P., 2005. Metadata: Automatic generation and extraction. In 7th MANLIBNET Annual National Convention on Digital Libraries in Knowledge Management: Opportunities for Management Libraries. Indian Institute of Management Kozhikode. http://dspace.iimk.ac.in/bitstream/2259/250/1/41-noufal-paper.pdf (Accedido por última vez en Febrero de 2011).

[33]. Mattmann, C.A., Zitting, J., 2011. Tika in Action, ISBN 13: 978-1-935182-85-6, Manning Early Access Program (MEAP).

[34]. Zong, W., Wu, D., Sun, A., Lim, E.P., Lian Goh, D.H., 2005. On Assigning Place Names to Geography Related Web Pages. Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries, Denver, CO, USA.

[35]. http://www.era.lib.ed.ac.uk/handle/1842/1849 ((Accedido por última vez en Agosto de 2011)

[36]. Jones, C. B., Purves, R. S., 2008. Geographical Information Retrieval International Journal of Geographical Information Science, 22, 219-228

[37]. Silva, M. J., Martins, B., Chaves, M., Afonso, A. P., Cardoso, N., 2006. Adding Geographic Scopes to Web Resources CEUS - Computers, Environment and Urban Systems, 30, 378-399

[38]. Campelo, C. E., Souza Baptista, C., 2009. A Model for Geographic Knowledge Extraction on Web Documents Proceedings of the ER 2009 Workshops (CoMoL, ETheCoM, FP-UML, MOST-ONISW, QoIS, RIGiM, SeCoGIS) on Advances in Conceptual Modeling - Challenging Perspectives, Springer-Verlag, 317-326

[39]. Inproceedings (Amitay2004) Amitay, E., Har'El, N., Sivan, R., Soffer, A., 2004. Web-a-Where: Geotagging Web Content SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 273-280

[40]. Jones, C. B., Alani, H., Tudhope, D., 2001. Geographical Information Retrieval with Ontologies of Place COSIT 2001: Proceedings of the International Conference on Spatial Information Theory, Springer-Verlag, 322-335

[41]. Overell, S., Rüger, S., 2008. Using co-occurrence models for placename disambiguation International Journal of Geographical Information Science, Taylor & Francis, Inc., 22, 265-287

[42]. Goldberg, D. W., 2008. A Geocoding Best Practices Guide North American Association of Central Cancer Registries (NAACCR), 287

[43]. Mens, T., Van Gorp, P., 2006. A taxonomy of model transformation, Electronic Notes in Theoretical Computer Science , vol. 152.

[44]. Biehl, M., 2010. Literature Study on Model Transformations. Royal Institute of Technology, Technical Report ISRN/KTH/MMK/R-10/07-SE. http://www.md.kth.se/~biehl/files/papers/mt.pdf (Accedido por última vez en Agosto de 2011).

[45]. Biehl, M., 2010. Literature Study on Model Transformations. Royal Institute of Technology, Technical Report ISRN/KTH/MMK/R-10/07-SE. http://www.md.kth.se/~biehl/files/papers/mt.pdf (Accedido por última vez en Agosto de 2011).

[46]. Czarnecki, K., Helsen, S., 2006. Feature-based survey of model transformation approaches. IBM Systems Journal, vol. 45, no. 3.

[47]. Open Geospatial Consortium, 2007. OpenGIS® Catalogue Services Specification, Reference number of this document: OGC 07-006r1, Version 2.0.2, Corrigendum 2 Release.

[48]. Cavnar, W.B., Trenkle, J.M., 1994. N-Gram-Based Text Categorization. In Proc. Third Annual Symposium on Document Analysis and Information Retrieval. UNLV.

[49]. Paynter, G., 2005. Developing Practical Automatic Metadata Assignment and Evaluation Tools for Internet Resources. Proc. Fifth ACM/IEEE Joint Conference on Digital Libraries, Denver, Colorado, June 7-11, ACM Press.

[50]. Humphreys, J.B., 2002. PhraseRate: An HTML Keyphrase Extractor. Technical report, University of California, Riverside.

[51]. Kedzierski, A., 2002. Artur's Auto Annotator. Masters Thesis, Department of Computer Science, University of California, Riverside.

[52]. Finkel, J.R., Grenager, T., Manning, C., 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005).

[53]. Abelson, H., Adida, B., Linksvayer, M., Yergler, N., 2008. ccREL: The Creative Commons Rights Expression Language. Technical report, Creative Commons. http://wiki.creativecommons.org/ images/d/d6/Ccrel-1.0.pdf (Accedido por última vez en Agosto de 2011)

[54]. Mitchell, S., Mooney, M., Mason J., Paynter G.W., Ruscheinski J., Kedzierski A., Humphreys K., 2003. iVia Open Source Virtual Library System. D-Lib Magazine 9, 1. January.

[55]. Borjas, B., Florczyk, A.J., Lopez-Pellicer, F.J., Nogueras-Iso, J., Zarazaga-Soria, F.J., 2011 Automatic Metadata Generation for the Web Geo-resources. INSPIRE Conference 2011: INSPIREd by 2020 - Contributing to smart, sustainable and inclusive growth. Edinburgh, Scotland, 27 June - 1 July 2011.

[56]. Borjas, B., Florczyk, A.J., López-Pellicer, F.J., Zarazaga-Soria, F.J., 2011. Generación Automática de Metadatos Geográficos de Páginas Web. 9th International Geomatics Week (Semana Geomática Internacional 2011). Barcelona, Spain, 15-17 March 2011.

[57]. Florczyk, F.J., López-Pellicer, A.J., Muro-Medrano, P.R., Nogueras-Iso, J., Zarazaga-Soria, F.J. 2010. Semantic Selection of Georeferencing Services for Urban Management. Journal of Information Technology in Construction. 2010, vol. 15 (Special Issue Bringing urban ontologies into practice). ISSN 0302-9743.

Anexo I. Acrónimos

API	*Application Programming Interface*
BBOX	*Bounding Box*
CSW	*Catalogue Services for the Web*
DC	*Dublin Core*
DCMI	*Dublin Core Metadata Iniciative*
EEES	Espacio Europeo de Educación Superior
GIR	*Geographic Information Retrieval*
GIS	*Geographic Information System*
GML	*Geography Markup Language*
GSDI	*Global Spatial Data Infrastructure Association*
GUID	*Globally Unique Identifier*
HTML	*HyperText Markup Language*
HTTP	*HyperText Transfer Protocol*
IAAA	Grupo de Sistemas de Información Avanzados
ICBM	*InterContinental Ballistic Missile*
IDE	Infraestructura de Datos Espacial
IDEE	Infraestructura de Datos Espaciales de España
INSPIRE	*Infrastructure for Spatial Information in Europe*
ISSO	*International Organization for Standardization*
KML	*Keyhole Markup Language*
MIME	*Multipurpose Internet Mail Extensions*
NER	*Name Entity Recognition*
OGC	*Open Geospatial Consortium*
OWS	*OGC Web Service*
POST	*Part-Of-Speech Tagging*
PURL	*Persistent Uniform Resource Locators*
RDF	*Resource Description Framework*
SAX	*Simple API for XML*
SIG	Sistema de Información Geográfica
SEO	*Search Engine Optimization*
TR	*Toponym Resolution*
URI	*Uniform Resource Identificator*
URL	*Uniform Resource Locator*
VGI	*Volunteered Geographic Information*
W3C	*World Wide Web Consortium*
WGS84	*World Geodetic System 1984*
WHATWG	*the Web Hypertext Application Technology Working Group*
WWW	*World Wide Web*
XHTML	*eXtensible Hypertext Markup Language*
XML	*eXtensible Markup Language*

Anexo II. Herramientas y Tecnologías

En esta anexo se incluyen las principales herramientas, tecnologías y especificaciones implicadas en las tareas de desarrollo de los diferentes componentes del proyecto.

Anexo II.I. Entorno de desarrollo

Java

Java[32] es una plataforma virtual de software desarrollada por *Sun Microsystems*, de manera que los programas creados en ella, puedan ejecutarse sin cambios en diferentes tipos de arquitecturas y dispositivos computacionales. La plataforma *Java* consta de las siguientes partes:

- El lenguaje de programación (*Java*).
- La máquina virtual de *Java* o JRE, que permite la portabilidad en ejecución.
- El API *Java*, una biblioteca estándar para el lenguaje.

Las principales características del lenguaje *Java* son:

- Es un lenguaje multiplataforma de propósito general.
- Usa la metodología de programación orientada a objetos.
- Posee mecanismos que permiten ejecutar aplicaciones remotas de forma segura.
- Existen numerosas utilidades y tecnologías disponibles.

Eclipse

Eclipse[33] es una plataforma o entorno de desarrollo de código fuente abierto y multiplataforma basado en el lenguaje *Java*. La característica más destacable de *Eclipse* es su extensibilidad, ya que es una gran estructura formada por un núcleo y múltiples complementos (*plugins)* que interactúan entre sí mediante interfaces o puntos de extensión, lo cual facilita la integración.

Está destinado para aplicaciones escritas en *Java*, pero es adaptable a cualquier otro lenguaje como C/C++, C#, XML, COBOL, etc.

Maven

Maven[34] es una herramienta de gestión de proyectos de software y automatización de su construcción. Se basa en un fichero de configuración, con formato XML, *Project Object Model* , POM, mediante el cual se pueden especificar los aspectos de construcción del proyecto, las dependencias con otros componentes y las acciones adicionales que se deseen realizar.

Maven realiza el tratamiento de las dependencias de forma recursiva, es decir, cuando se requiere construir un proyecto con determinadas dependencias a otros componentes, lo cuales a su vez tienen otras dependencias, estas últimas se incluyen automáticamente en el proyecto inicial sin tener que volverlas a especificar. De este modo se consigue simplificar enormemente la tarea de uso de las librerías externas.

Una de las ventajas más significativas de *Maven* es la utilización de repositorios de los que se obtienen los componentes especificados en las dependencias. Permite obtener automáticamente las librerías de los repositorios indicados, ya sean remotos o locales.

[32] http://java.sun.com/
[33] http://www.eclipse.org/
[34] http://maven.apache.org/

Un hecho a destacar es la existencia de complementos para integrar *Maven* con el entorno de desarrollo *Eclipse* o añadir otras funcionalidades en el proceso de construcción del proyecto (generación de documentación *javadoc*, informes de ejecución de pruebas (*tests*), comprobación del formato del código fuente, etc.).

Subversion

Subversion[35] (conocido también como SVN) es un sistema de control de versiones de código fuente abierto. Es decir, *Subversion* gestiona ficheros y directorios a través del tiempo. Hay un árbol de ficheros en un repositorio central. El repositorio es como un servidor de ficheros ordinario, excepto porque recuerda todos los cambios hechos a sus ficheros y directorios. Esto le permite recuperar versiones antiguas de sus datos, o examinar el historial de cambios de los mismos.

Una característica importante de *Subversion* es que, los ficheros versionados no tienen cada uno un número de revisión independiente, en cambio, todo el repositorio tiene un único número de versión que identifica un estado común de todos los ficheros del repositorio en un instante determinado.

Existe un complemento denominado *Subclipse*[36] que permite integrar *Subversion* con el entorno de desarrollo *Eclipse*.

Anexo II.II. Estándares y especificaciones

ISO 19115

La norma ISO 19115 (ISO 19115:2003 *Geographic Information Metadata*) proporciona un modelo o esquema y establece un conjunto común de terminología, definiciones y procedimientos de aplicación para los elementos de metadatos que describen la información geográfica. Esta norma provee de información sobre la identificación, la extensión, la calidad, el modelo espacial y temporal, la referencia espacial y la distribución de los datos geográficos digitales.

Aunque la norma ISO 19115, que es de gran complejidad, define un extenso número de elementos de metadatos, establece un conjunto mínimo de ellos a considerar. Con este conjunto se pretende establecer unos mínimos para facilitar el descubrimiento, el acceso, la transferencia y la utilización de los datos digitales. A pesar de que esta norma es aplicable a los datos digitales, sus principios pueden extenderse a muchas otras formas de datos geográficos tales como mapas, cartas y documentos de texto, así como los datos no geográficos.

La norma ISO 19115 proporciona una estructura para describir información geográfica mediante elementos de metadatos pero no desarrolla cómo poder llevar a cabo su implementación. La norma ISO 19139 (ISO/TS 19139:2007*Geographic Information-Metadata -XML schema implementation*) es una especificación técnica que desarrolla una implementación en el lenguaje de marcado XML del modelo de metadatos descrito por ISO 19115.

Dublin Core

La iniciativa de metadatos *Dublin Core*, *Dublin Core Metadata Iniciative* (DCMI)[37], promueve la difusión de los estándares o normas de metadatos interoperables y el desarrollo de vocabularios

[35] http://subversion.apache.org/
[36] http://subclipse.tigris.org/
[37] http://dublincore.org/documents/dcmi-terms/

de metadatos especializados que permiten la construcción de sistemas de búsqueda de información más inteligentes.

Dublin Core es una norma para la descripción de todo tipo de recursos independientemente de su formato, área de especialización u origen cultural. Tiene carácter oficial, ya que se ha aprobado como norma americana (ANSI/NISO Z39.85 - 2007 *The Dublin Core Metadata Element Set*)[38], se ha adaptado dentro del comité técnico europeo (CEN/ISSS *Workshop on Meta-Data (Dublin Core)*)[39], y también tiene carácter de norma internacional ISO (ISO 15836:2009 *Information and documentation The Dublin Core metadata element set*).

Esta norma consiste en quince descriptores básicos que son el resultado de un consenso internacional e interdisciplinario[40]. La simplicidad de *Dublin Core* permite un fácil emparejamiento con otros esquemas de metadatos más específicos, lo que hace que muchas organizaciones consideren la adopción de *Dublin Core* en determinadas situaciones. En el ámbito de los SIG se considera la adopción de *Dublin Core* en determinadas situaciones. De hecho, la especificación de los servicios de catálogo para recursos Web, propuesta por la OGC, propone usar *Dublin Core* como modelo básico de búsqueda y presentación de metadatos para la descripción de los recursos geográficos.

Catalog Service for Web

La especificación de la OGC, Servicios de Catálogo para los recursos Web, *Catalog Service for Web*, CSW (OGC *Catalogue Services Implementation Specificacion* 2.0.2:2007), describe un conjunto de interfaces de las operaciones que soportan la gestión, el descubrimiento, y el acceso a los recursos de información geográfica. La gestión proporciona la funcionalidad de organización de las entradas del catálogo en el dispositivo de almacenamiento local (por ejemplo, sistemas de ficheros o bases de datos relacionales). El descubrimiento permite que los usuarios busquen dentro del catálogo usando un lenguaje de consulta con una sintaxis reconocida. Las consultas se proponen en el lenguaje denominado OGC *Common Query Language*, muy similar al utilizado en las cláusulas *WHERE* de SQL. Además, se propone la codificación de este lenguaje sobre XML utilizando la especificación *Filter Encoding Specification*. El servicio de acceso facilita la interacción con los elementos que se habían previamente localizado con los servicios de descubrimiento.

Un aspecto importante en esta especificación es que se proporcionan diferentes perfiles de implementación de las interfaces de acuerdo a la plataforma y protocolo de transporte (*protocol binding*) que se va a utilizar. Tomando como ejemplo el protocolo CSW, las operaciones que un catálogo ofrece son las siguientes:

- *GetCapabilities*: Operación que obtiene los metadatos que describen de forma general los contenidos y las capacidades de un servidor particular que está implementando la especificación de los servicios de catálogo para acceder a una colección concreta de metadatos. La respuesta a una petición *GetCapabilities* será un documento con una codificación basada en XML, denominado Registro CSW, que contiene los metadatos del servicio y que están expresados usando la sintaxis y la nomenclatura propuesta por la iniciativa de metadatos *Dublin Core* (ISO 15836). En la Tabla 9 se muestra un mapeo entre los nombres de elementos de *Dublin Core*, los principales términos de consultas OGC y los elementos concretos de XML.

[38] http://www.niso.org/apps/group_public/project/details.php?project_id=57
[39] http://www.cen.eu/cen/Sectors/Sectors/ISSS/Activity/Pages/WSMMI.aspx
[40] http://dublincore.org/documents/dces/

Dublin Core	Término OGC	Elemento XML
contributor		`dc:contributor`
coverage	*BoundingBox*	`ows:BoundingBox`
creator		`dc:creator`
date	*Modified*	`dct:modified`
description	*Abstract*	`dct:abstract`
format	*Format*	`dc:format`
identifier	*Identifier*	`dc:identifier`
language		`dc:language`
publisher		`dc:publisher`
relation	*Association*	`dc:relation`
rights		`dc:Rights`
source	*Source*	`dc:source`
subject	*Subject*	`dc:subject`
title	*Title*	`dc:title`
type	*Type*	`dc:type`

Tabla 9: El mapeo entre nombres Dublin Core y nombres de elementos XML[41]

- *DescribeRecord*: Operación de descubrimiento que permite que el cliente descubra el modelo de información de los metadatos ofrecidos a través del servidor de catálogo.
- *GetDomain*: Operación de descubrimiento de metadatos que devuelve el rango de valores de las propiedades consultadas en tiempo de ejecución.
- *GetRecords*: Operación de descubrimiento de devuelve el conjunto de registros de metadatos que cumplen las restricciones de las consultas especificadas por un cliente.
- *GetRecordById*: Operación de descubrimiento de metadatos que devuelve un registro con un identificador determinado.
- *Transaction*: Operación de gestión que permite realizar la inserción, borrado y modificación de un registro en el catálogo.
- *Harvest*: Operación de gestión que realiza la actualización o modificación de registros de una forma asíncrona.

[41] El esquema: http://schemas.opengis.net/csw/2.0.0/rec-dcterms.xsd contiene una lista completa de elementos XML sustituibles.

Anexo II.III. Metadatos utilizados frecuentemente

Metatags.org

El sitio web comercial Metatags.org[42] es gestionado por la empresa holandesa *The Metatags Company Inc.* que opera en diferentes países y provee servicios de mercadotecnia para ayudar a las empresas a optimizar su presencia en la red y su posicionamiento en los diferentes motores de búsqueda. Según este sitio Web, la investigación muestra que sólo el 20% de todas las páginas Web contienen elementos `META` en su código HTML y más del 85% de estos sitios Web no son aptos para ser presentados a los motores de búsqueda. En la llamada SEO, la optimización en los motores de búsqueda (*Search Engine Optimization*), los elementos `META` todavía juegan un papel importante. Especialmente el uso del elemento de descripción (*description*) es muy importante ya que éste se muestra por los motores de búsqueda dentro de un pequeño recuadro de texto en los resultados de búsqueda. En la tabla 10 se listan algunos de los nombres de metadatos utilizados frecuentemente en la Web y su influencia SEO de acuerdo al mencionado sitio Web.

Metadato	Elemento HTML	Descripción	Influencia SEO
abstract	`META` (*name*)	Descripción muy corta	Baja
author	`META` (*name*)	Autor del contenido	Ninguna
contact	`META` (*name*)	Persona de contacto	Ninguna
content-type	`META` (*equiv*)	Tipo de medio y conjunto de caracteres	Ninguna
copyright	`META` (*name*)	Licencia de copia	
creation_date	`META` (*name*)	Fecha de creación	Ninguna
*dc.**	`META` (*name*)	*Dublin Core Metadata Initiative*	Baja
description	`META` (*name*)	Descripción	Alta
distribution	`META` (*name*)	Nivel de distribución	Baja
expires	`META` (*name*)	Expiración del contenido	Baja
generator	`META` (*name*)	Programa generador	
identifier-URL	`META` (*name*)	URL	Ninguna
keywords	`META` (*name*)	Palabras claves	Baja
language	`META` (*name*)	Lenguaje	
rating	`META` (*name*)	Adecuación a la audiencia	Baja
refresh	`META` (*equiv*)	Período de refresco o redirección	Alta
resource-type	`META` (*equiv*)	Tipo de recurso	Ninguna
revisit-after	`META` (*name*)	Período visita de los *crawlers*	
robots	`META` (*name*)	Permitir acceso a Robots	
subject	`META` (*name*)	Tema	Ninguna
title	`TITLE`	Título	Alta
webauthor	`META` (*name*)	Compañía que desarrolla el sitio	

Tabla 10: Metadatos usados frecuentemente en la Web según Metatags.org

[42] http://www.metatags.org/

HTML 5 - *MetaExtensions*

En la quinta revisión de W3C del lenguaje básico de la Web, HTML 5[43], se especifica por vez primera un conjunto de algunos nombres de forma estándar para el atributo `NAME` del elemento `META`. Los nombres de metadatos son insensibles a las mayúsculas y minúsculas. Las extensiones para este conjunto predefinido de nombres se pueden proponer y registrar en la página *Wiki* de *MetaExtensions*[44] gestionado por *The Web Hypertext Application Technology Working Group*, WHATWG[45]. Los nombres de metadatos cuyos valores sean URL no podrán ser propuestos o aceptados, ya que éstos se representan mediante el elemento `LINK`.

En la tabla 11 se describen los nombres estándares de metadatos especificados por W3C y las principales extensiones de nombres propuestos en el sitio *MetaExtensions*.

Metadato	Descripción	Estatus
application-name	Nombre de la aplicación que la página representa	Estándar
Audience	Audiencia mas apropiada para la página	Propuesta incompleta
Author	Nombre de cada uno de los autores de la página	Estándar
Creador	Creador o autor de la página	Propuesta incompleta
DC.	Prefijo para todos los elementos de *Dublin Core MetaData Initiative*	Propuesta incompleta
dc.language	Lenguaje del recurso (vocabulario controlado)	Propuesta
dcterms.abstract	Resumen del recurso	Propuesta
dcterms.accessRights	Información sobre derechos de acceso al recurso o estatus de seguridad	Propuesta
dcterms.alternative	Nombre alternativo para el recurso	Propuesta
dcterms.audience	Clase de entidad para el cual el recurso está intencionado o es útil	Propuesta
dcterms.available	Fecha de disponibilidad del recurso	Propuesta
dcterms.bibliographic Citation	Referencia bibliográfica del recurso	Propuesta
dcterms.contibutor	Entidad responsable de hacer contribuciones al recurso	Propuesta
dcterms.coverage	Tópico espacial o temporal del recurso	Propuesta
dcterms.creates	Fecha de creación del recurso	Propuesta
dcterms.creator	La entidad encargada principalmente de crear el recurso	Propuesta
dcterms.date	Periodo de tiempo asociado a evento en ciclo de vida del recurso	Propuesta
dcterms.extent	Tamaño o duración del recurso	Propuesta
dcterms.format	Formato de fichero, medio físico o dimensiones del recurso	Propuesta
dcterms.identifier	Referencia no ambigua del recurso dentro de un contexto dado	Propuesta
dcterms.isReferenceBy	Recurso relacionado que hace referencia o cita el recurso descrito	Propuesta
dcterms.language	Lenguaje del recurso (vocabulario controlado)	Propuesta
dcterms.license	Un documento legal que da permiso oficial para usar el recurso	Propuesta
dcterms.publisher	La entidad responsable de poner disponibles los recursos	Propuesta
dcterms.references	Recurso relacionado que es referenciado o citado	Propuesta

[43] http://dev.w3.org/html5/spec/Overview.html#the-meta-element
[44] http://wiki.whatwg.org/wiki/MetaExtensions
[45] http://www.whatwg.org/

dcterms.relation	Recurso relacionado	Propuesta
dcterms.rights	Información sobre los derechos acerca del recurso	Propuesta
dcterms.rightsHolder	Persona u organización propietaria o gestora de los derechos del recurso	Propuesta
dcterms.source	Recurso relacionado del cual se deriva el recurso descrito	Propuesta
dcterms.spatial	Características espaciales del recurso	Propuesta
dcterms.subject	Temática del recurso	Propuesta
dcterms.tableOfContens	Lista de subunidades del recurso	Propuesta
dcterms.type	Naturaleza o género del recurso	Propuesta
dcterms.valid	Fecha o rango de fechas de validez del recurso	Propuesta
description	Descripción de la página	Estándar
Designer	Créditos al diseñador responsable de la presentación visual	Propuesta
Expires	Fecha de expiración de la página	Propuesta Incompleta
Generador	Paquete de software usado para generar la página	Estándar
geo.country	Código de un país con el que está relacionada la página	Propuesta
geo.placename	Nombre del lugar geográfico con el que está relacionada la página	Propuesta
geo.position	Posición geográfica con el que está relacionada la página	Propuesta
geo.region	Código de región geográfica con el que está relacionada la página	Propuesta
geographic-coverage	Información sobre extensión geográfica (lista de lugares)	Propuesta Incompleta
ICBM	Posición geográfica con el que está relacionada la página	Propuesta
keywords	Palabras claves relevantes para la página	Estándar
Publisher	Entidad responsable de la publicación del contenido	Propuesta
Rating	Restricción de contenido para adultos (RTA)	Propuesta
review_date	Fecha programada para la revisión del recurso	Propuesta
Rights	Derechos de propiedad intelectual del recurso	Propuesta Incompleta
rights-standard	Tipos de los derechos asignados a la obra	Propuesta
Robots	Lista de operadores para explicar tratamiento de contenido a *crawlers*	Propuesta

Tabla 11: Nombres estándar y propuestos de metadatos según W3C

Anexo III. Geoportales

Las siguientes tablas a continuación recogen información de las diversas iniciativas, de ámbito global, regional, nacional y local, relacionadas con las infraestructuras de datos espaciales, recopiladas por la *Global Spatial Data Infrastructure Association*, GSDI[46]. Esta asociación es una organización inclusiva de organizaciones, agencias, empresas e individuos de todo el mundo El propósito de la organización es promover la cooperación y colaboración internacional en apoyo de las infraestructuras de datos espaciales locales, nacionales e internacionales, que permitan a las naciones abordar mejor las cuestiones sociales, económicas y ambientales que tengan una importancia apremiante.

Cód.	URL	OK	Descripción	Ámbito
G01	http://www.csi-cgiar.org	✓	Consortium for Spatial Information	Global
G02	http://www.cgiar.org/	✓	Consultative Group on International Agriculture (CGIAR)	Global
G03	http://www.digitalearth.gov		Digital Earth	Global
G04	http://earthexplorer.usgs.gov	✓	Earth Explorer	Global
G05	http://www.fao.org/sd/EIdirect/gis/EIgis000.htm	✓	Geographic Information Systems in Sustainable Development	Global
G06	http://www.gdin.org	✓	Global Disaster Information Network (GDIN)	Global
G07	http://glcf.umiacs.umd.edu/index.shtml	✓	Global Land Cover Facility	Global
G08	http://www.iscgm.org/cgi-bin/fswiki/wiki.cgi	✓	Global Map	Global
G09	http://www.grid.unep.ch/	✓	Global Resource Information Database (GRID)	Global
G10	http://grdc.bafg.de/servlet/is/2479/	✓	Global Water Information Network (GLOBWINET)	Global
G11	http://www.ipcc.ch/index.htm	✓	Intergovernmental Panel on Climate Change (IPCC)	Global
G12	http://www.icao.int/	✓	International Civil Aviation Organization (ICAO)	Global
G13	http://www.igbp.kva.se/	✓	International Geosphere - Biosphere Program (IGBP)	Global
G14	http://www.isotc211.org	✓	International Organization for Standardization of Geographic information/Geomatics ISO/TC 211	Global
G15	http://www.natureserve.org	✓	NatureServe	Global
G16	http://www.opengeospatial.org	✓	Open Geospatial Consortium (OGC)	Global
G17	http://srtm.usgs.gov	✓	Shuttle Radar Topography Mission	Global
G18	http://www.ungiwg.org/	✓	United Nations Geographic Information Working Group (UNGIWG)	Global
G19	http://www.unrisd.org	✓	United Nations Research Institute for Social Development	Global
G20	http://earthwatch.unep.net		United Nations System-Wide Earthwatch	Global
G21	http://www.developmentgateway.org	✓	World Bank Development Gateway	Global

Tabla 12: IDEs: Iniciativas de ámbito global (último acceso: octubre 2011)

[46] http://www.gsdi.org/SDILinks

Cód.	URL	OK	Descripción	Ámbito
R01	http://www.uneca.org/awich	✓	African Regional Water Resources Information System (AWICH)	Africa
R02	http://www.antsdi.scar.org		Antarctic Spatial Data Infrastructure.	Antarctica
R03	http://www.ANZLIC.ORG.au/	✓	ANZLIC - The Spatial Information Council	Australia/ New Zealand
R04	http://www.arcus.org/gis/forum.html	✓	Arctic GIS	Arctic
R05	http://www.caribbeanGIS.com	✓	Caribbean GIS	Caribbean
R06	http://www.cep.unep.org/search/search.htm#"	✓	Caribbean Regional Co-ordinating Unit-Spatial Data Clearinghouse, UNEP	Caribbean
R07	http://www.PROCIG.org	✓	Central American Geographic Information Project (PROCIG)	Central America
R08	http://www.MAPBSR.NLS.fi		Digital Map of The Baltic Sea Region. Baltic Sea Region	
R09	http://www.dnf.org	✓	Digital National Framework (DNF)	UK
R10	http://www.EIS-Africa.org/	✓	Environmental Information Systems-Africa (EIS Africa)	Africa
R11	http://www.eurogeoraphics.org/eng/01_about.php		Europe's National Mapping Agencies (EuroGeoGraphics)	Europe
R12	http://eusoils.jrc.it	✓	European Soil Portal	Europe
R13	http://www.ec-gis.org/etemii/	✓	European Territorial Management Information Infrastructure	Europe
R14	http://www.EUROGI.org	✓	European Umbrella Organisation for Geographic Information (EUROGI)	Europe
R15	http://www.uneca.org/disd/geoinfo/	✓	Geo Information, United Nations Economic Commission for Africa, Information Services	Africa
R16	http://www.state.gov/g/oes/rls/fs/2002/15618.htm		Geographic Information for Sustainable Development in Africa	Africa
R17	http://www.ec-gis.org/ginie/	✓	Geographic Information Network in Europe (GINIE)	Europe
R18	http://www.gisdevelopment.net/policy/gi/gii0008.htm		ICIMOD's Approach Towards A Regional Geo-Information Infrastructure (RGII) In The Hindu-Kush	Himalayan (HKH) Region
R19	http://www.ec-gis.org/inspire/		Infrastructure for Spatial Information in Europe (INSPIRE)	Europe
R20	http://www.PCGIAP.org	✓	Permanent Committee on GIS Infrastructure for Asia & The Pacific (PCGIAP)	Asia and Pacific
R21	http://www.CPIDEA.org	✓	Permanent Committee on SDI for The Americas (PC IDEA)	The Americas
R22	http://www.uneca.org/disd/ict/index.htm	✓	United Nations Economic Commission for Africa	Africa
R17	http://www.ec-gis.org/ginie/	✓	Geographic Information Network in Europe (GINIE)	Europe
R18	http://www.gisdevelopment.net/policy/gi/gii0008.htm	✓	ICIMOD's Approach Towards A Regional Geo-Information Infrastructure (RGII) In The Hindu-Kush	Himalayan (HKH) Region
R19	http://www.ec-gis.org/inspire/	✓	Infrastructure for Spatial Information in Europe (INSPIRE)	Europe
R20	http://www.PCGIAP.org	✓	Permanent Committee on GIS Infra-structure for Asia & Pacific (PCGIAP)	Asia and Pacific
R21	http://www.CPIDEA.org	✓	Permanent Committee on SDI for The Americas (PC IDEA)	The Americas
R22	http://www.uneca.org/disd/ict/index.ht	✓	United Nations Economic Commission	Africa

Tabla 13: IDE's: Iniciativas de ámbito regional (último acceso: octubre 2011)

Cód.	URL	OK	Descripción	Ámbito
N01	http://www.AMFM.it		Am/Fm Geographic Information System	Italy
N02	http://www.osdm.gov.au/osdm/spatial_data.html	✓	Australian Government Spatial Data	Australia
N03	http://asdd.ga.gov.au/asdd/	✓	Australian Spatial Data Directory	Australia
N04	http://www.CGDI.GC.CA	✓	Canadian Geospatial Data Infrastructure (CGDI)	Canada
N05	http://CLEARINGHOUSE.CNR.gob.sv/		Clearinghouse de El Salvador	Salvador
N06	http://www.CLEARINGHOUSE.gob.ni/		Clearinghouse de Nicaragua	Nicaragua
N07	http://www.clearinghouse.com.uy%20		Clearinghouse Nacional de Datos Geográficos	Uruguay
N08	http://www.cnig.gouv.fr/	✓	Conseil National de l'Information Géographique	France
N09	http://www.CAGI.cz	✓	Czech Association for Geoinformation (CAGI)	Czech Republic
N10	http://www.FCCLAND.ru/	✓	FCC Zemlya	Russia
N11	http://www.GEOFORUM.no/	✓	Geoforum	Norway
N12	http://www.GEOFORUM.dk/	✓	Geoforum	Denmark
N13	http://www.mosaic-ni.gov.uk/		Geographic Information Strategy for Northern Ireland	Northern Ireland
N14	http://www.DDGI.de	✓	German Umbrella Organization for Geoinformation	German
N15	http://www.gigateway.org.uk/		GI Gateway	United Kingdom
N16	http://www.SIGOV.si/	✓	GI Portal	Slovenia Republic
N17	http://www.MLMUPC.gov.kh	✓	GIS Task Force	Kingdom of Cambodia
N18	http://www.GISPOL.ORG.pl/		GISPOL	Poland
N19	http://www.fomi.hu/hunagi/	✓	Hungarian Association for Geo-Information	Hungary
N20	http://www.idee.es	✓	IDEE for Infraestructura de Datos Espaciales de España	Spain
N21	http://www.ICDE.ORG.co	✓	Infraestructura Colombiana de Datos Espaciales ICDE	Colombia
N22	http://www.IGM.cl	✓	Infraestructura Nacional de Datos Espaciales (INDE)	Chile
N23	http://www.IRLOGI.ie	✓	Irish Organisation for Geographic Information IRLOGI	Ireland
N24	http://www.NALIS.gov.my http://www.mygeoportal.gov.my	✓	Malaysia Geospatial Data Infrastucture (MyGDI)	Malaysia
N25	http://www.bakosurtanal.go.id/	✓	National Coordination Agency for Surveys and Mapping	Indonesia
N26	http://NFGIS.NSDI.gov.cn	✓	National Fundamental Geographic Information System of China	China
N27	http://www.nls.fi/ptk/infrastructure/	✓	National Geographic Information	Finland
N28	http://gcmd.nasa.gov/records/NSDI_INDIA.html		National Spatial Data Infrastructure(NSDI)	India
N29	http://www.FGDC.gov/	✓	National Spatial Data Infrastructure (NSDI)	United States

N30	http://www.NSDIPA.gr.jp/	✓	National Spatial Data Infrastructure Promoting Association	Japan
N31	http://www.NSIF.ORG.za/		National Spatial Information Framework	South Africa
N32	http://www.RVK.IS/lisa	✓	Organisation of Geographical Information for All In Iceland –LISA	Iceland
N33	http://www.SOGI.ch/	✓	Organisation Suisse Pour l'Information Geographique	Switzerland
N34	http://www.psdn.org.ph/iatfgi/ngii.htm		Philippines National Geographic Information Infrastructure (NGII)	Philippines
N35	http://www.RAVI.nl/		RAVI	The Netherlands
N36	http://snig.igeo.pt/	✓	Sistema Nacional De Informação Geográfica (SNIG)	Portugal
N37	http://www.CMW.INF.cu		Spatial Data Infrastructure of the Cuban Republic –IDERC	Cuba
N38	http://www.gisig.it/	✓	Spatial Data Infrastructures in S.E.Europe	Italy

Tabla 14: IDEs - Iniciativas de ámbito nacional (último acceso: octubre 2011)

Cód.	URL	OK	Descripción	Ámbito
L01	http://www.ntlis.nt.gov.au	✓	Northern Territory Land Information System (NTLIS)	Northern Australia
L02	http://www.qsiis.qld.gov.au/		Queensland Spatial Information Infrastructure Strategy (QSIIS)	Queensland Australia
L03	http://services.land.vic.gov.au/landchannel/content/productCatalogue	✓	Vicmap	Victoria, Australia
L04	http://www.walis.wa.gov.au/		Western Australian Land Information System (WALIS)	Western Australia
L05	http://www.ideandalucia.es/	✓	SDI of Andalucía (IDEAndalucía)	Andalucía, Spain
L06	http://sitar.aragon.es/	✓	SDI of Aragon (SITAR)	Aragon, Spain
L07	http://www.idecan.grafcan.es/idecan/	✓	SDI of Canarias (IDECanarias)	Canarias, Spain
L08	http://ide.jccm.es/	✓	SDI of Castilla-La Mancha (IDEclm)	Castilla-La Mancha, Spain
L09	http://www.sitcyl.jcyl.es/sitcyl/home.sit	✓	SDI of Castilla y León (IDECyL)	Castilla y León, Spain
L10	http://www.geoportal-idec.net/geoportal/cat/inici.jsp	✓	SDI of Cataluña (IDEC)	Cataluña, Spain
L11	http://idena.navarra.es/busquedas/catalog/main/home.page	✓	SDI of Navarra (IDENA)	Navarra, Spain
L12	http://www.icv.gva.es/es/	✓	SDI of Valencia (IDECV)	Valencia, Spain
L13	http://www.ideextremadura.es/	✓	SDI of Extremadura (IDEextremadura)	Extrema-dura, Spain
L14	http://www.ideib.cat/index.php?newlang=spanish	✓	SDI of Baleares (IDEIB)	Baleares, Spain
L15	http://sitga.xunta.es/sitganet/	✓	SDI of Galicia (SITGA)	Galicia, Spain
L16	http://www.iderioja.larioja.org/	✓	SDI of La Rioja (IDErioja)	La Rioja, Spain
L17	http://www.geo.euskadi.net/s69-8241/es/	✓	SDI of País Vasco (GeoEuskadi)	País Vasco, Spain

L18	http://www.cartografia.princast.es/cartositpa/	✓	SDI of Asturias (SitpaIdeas)	Asturias, Spain
L19	http://www.cartomur.com/	✓	SDI of Murcia (Cartomur)	Murcia, Spain
L20	http://www.thelist.tas.gov.au/		Land Information System Tasmania (The LIST)	Tasmania
L21	http://portal.gsa.state.al.us/Portal/index.jsp	✓		Alabama, USA
L22	http://www.asgdc.state.ak.us/	✓	Alaska State Geo-spatial Data Clearinghouse (ASGDC)	Alaska, USA
L23	http://www.geostor.arkansas.gov	✓		Arkansas, USA
L24	http://gis.ca.gov/catalog/	✓		California, USA
L25	http://coloradogis.nsm.du.edu	✓		Colorado, USA
L26	http://www.ct.gov/gis	✓		Connecticut, USA
L27	http://gis.smith.udel.edu/fgdc/gateway/			USA
L28	http://data.georgiaspatial.org	✓		Georgia, USA
L29	http://www.insideidaho.org	✓		Idaho, USA
L30	http://www.isgs.uiuc.edu/nsdihome	✓		Illinois, USA
L31	http://indianamap.org	✓		Indiana, USA
L32	http://www.iowagis.org	✓		Iowa,
L33	http://www.kansasgis.org	✓		Kansas, USA
L34	http://kygeonet.ky.gov/	✓		Kentucky, USA
L35	http://doa.louisiana.gov/lgisc/	✓		Louisiana, USA
L36	http://megis.maine.gov	✓		Maine, USA
L37	http://www.marylandgis.net	✓		Maryland, USA
L38	http://www.michigan.gov/csstp	✓		Michigan, USA
L39	http://www.lmic.state.mn.us/chouse/index.html	✓		Minnesota, USA
L40	http://wgiac2.state.wy.us/	✓	Wyoming Spatial Data Clearinghouse	Wyoming, USA
L41	http://www.csc.noaa.gov/data/	✓	Coastal NSDI	Charleston, SC, USA
L42	http://www.sdvc.uwyo.edu/gya/	✓	Yellowstone National Spatial Data Infrastructure Initiative	Yellowstone, USA

Tabla 15: IDE's - Iniciativas de ámbito local (último acceso: octubre 2011)

Anexo IV. Diseño Detallado

Modelo de interfaces

El diagrama mostrado en la figura 6 describe el modelo que representa las interfaces en el cual se basa el diseño del prototipo desarrollado para la implementación de la arquitectura propuesta de caracterización automática de metadatos para recursos Web.

La clase *Metadata* (`org.apache.tika`) es la responsable de la gestión (creación, consulta o eliminación) de las propiedades o atributos multivaluados de un contenedor de metadatos. La clase *MetadataDocument* permite obtener la localización URL del documento de entrada y determinar su esquema, y la clase *InputStreamMetadataCreator* se encarga de crear la ráfaga (*stream*) de *bytes* de entrada del recurso referenciado en el documento de entrada. Además, opcionalmente se genera un conjunto de metadatos de entrada que contiene la información obtenida al conectar con el recurso. Las constantes para estas propiedades, propias del recurso, están declaradas en las clases *FileMetadata*, *MagicTypeDetectionMetadata* y *ResourceMetadata*. La clase *Criterion* es usada para pasar u obtener información de contexto entre los diferentes componentes del sistema.

La clase *MetadataGeneratorProvider* es la responsable de la generación de los metadatos del recurso que viene referenciado desde el documento de entrada. El documento de entrada podría ser de diferentes esquemas (estándares o vocabularios). Las tareas de esta clase son: 1) detectar el esquema del documento de entrada; 2) localizar el recurso cuya propiedades de metadatos deberían ser aumentadas; 3) cargar el extractor o los extractores, de acuerdo a un fichero de configuración; 4) extraer el conjunto (o conjuntos) de metadatos para incrementar la información original de los metadatos del recurso referenciado en documento de entrada.

La clase *MetadataHolder* permite asignar u obtener la instancia asociada de la clase *Metadata* y la clase *MetadataSupporter* permite la verificación de la existencia de soporte para el tipo MIME y la verificación del tipo o clase del conjunto de metadatos.

La clase *ExtractorManager* es la responsable de la creación de la instancia de una clase de extractores y la clase *MetadataExtractor* de la extracción del conjunto de metadatos. La clase *ExtractorInfo* permite obtener información acerca de las propiedades o características de un extractor. La clase *MetadataExtractorProvider* ofrece una forma configurable para la extracción de los metadatos. A través de un fichero de configuración se podría informar de las características por defecto o las propiedades (por ejemplo los tipos MIME soportados o la codificación de la ráfaga (*stream*) de bytes de la entrada del recurso). En el caso de XML, el extractor trabaja como un analizador sintáctico (*parser*) y en la configuración deberá estar registrada la clase controladora (*handler*) que usa el API basada en eventos, SAX, *Simple API for XML*, que deberá ser instanciada para el procesamiento de un documento XML.

La clase *HTMLMapper* permite realizar un mapeo de nombres de elementos y atributos HTML "seguros" con su equivalente semántico en XHTML. Si el elemento se desconoce o se considera nocivo para su inclusión en el análisis, el elemento será ignorado, aunque su contenido podrá seguir siendo procesado. En el caso de un atributo desconocido, éste se ignora. La clase *ContentHandler* (`org.xml.sax`) recibe la notificación del contenido lógico de un documento. La clase *MetaHtmlHandler* gestiona los eventos SAX para la extracción de contenido del documento del recurso XHTML y la asignación de propiedades o atributos y sus valores dentro de un contenedor de metadatos.

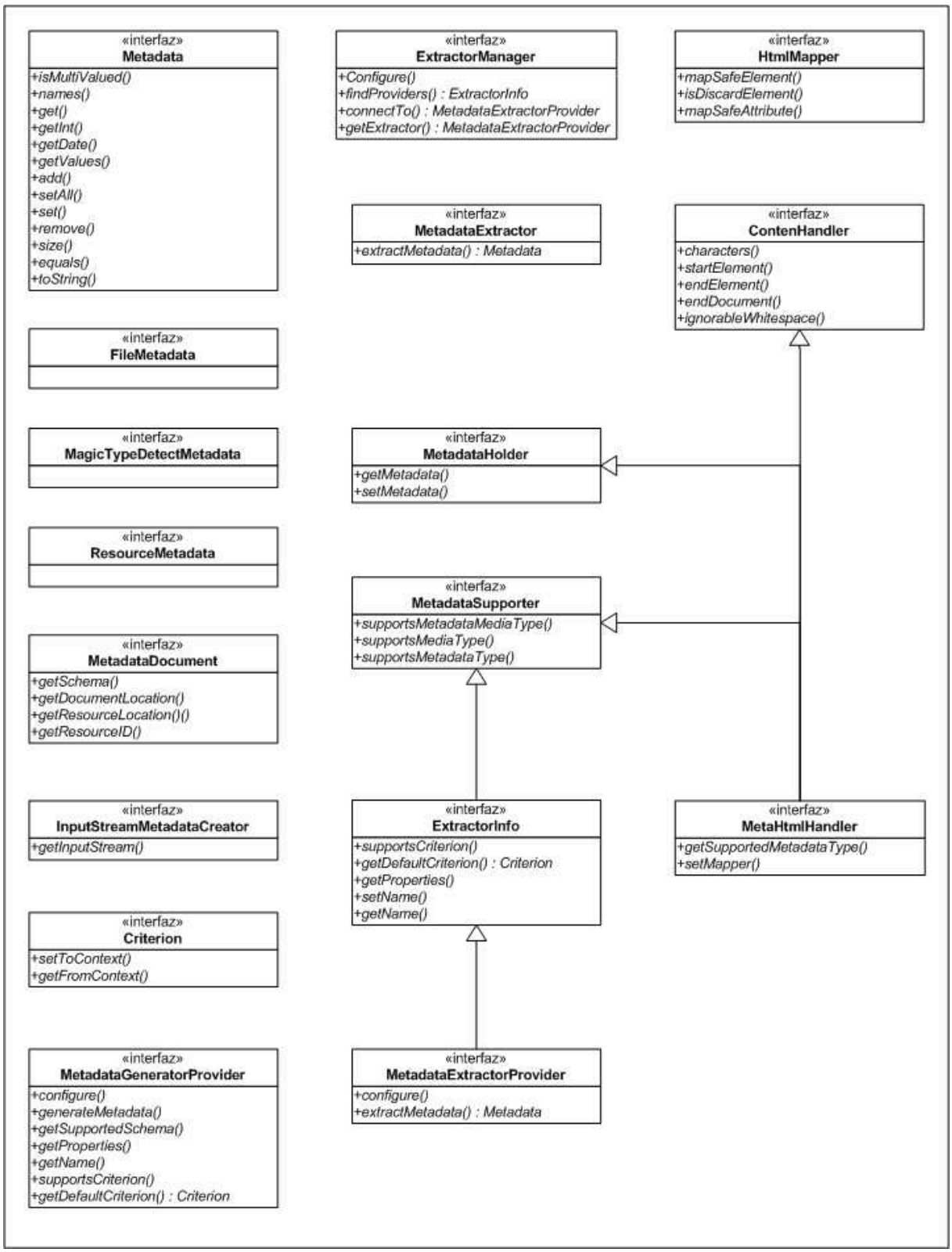

Figura 6: Modelo de interfaces

Anexo V. Desarrollo de Heurísticas para Estimación de Extensión Geográfica

Anexo V.I. Análisis Manual

En la tabla 16 se presentan los resultados del análisis manual de una muestra de páginas principales de geoportales para el desarrollo de heurísticas para la estimación de la extensión geográfica. Se han analizado los siguientes elementos de la página Web: (1) Posición (longitud/latitud) y nombre geográfico de elementos `META` (geográficos), (2) nombres geográficos de elementos `META` (no geográficos) y `TITLE`, (3) nombres geográficos del texto visible de los enlaces (`A`) del elemento `BODY`, 4) nombres geográficos del texto no visible (`alt`) de las imágenes (`IMG`) del elemento `BODY`, (5) nombres geográficos del texto visible del elemento `BODY` sin 3) y sin 4).

En el caso de los geoportales globales se ha observado que la fuente principal de los nombres geográficos es el (5). Además, hay mucha variedad de nombres de países que pertenecen a las regiones de diferentes partes del mundo.

En el caso de los geoportales regionales se ha observado que (3) y (5) son la fuente principal de los nombres geográficos y los nombres de países pertenecen a una región geográfica.

En el caso de los geoportales nacionales, la principal característica es la escasez de los nombres geográficos identificados en (3), (4) y (5). Suele aparecer sólo un nombre de país, eventualmente el nombre de su capital. El análisis también identificó en este caso que el código de país que aparece en el dominio del *source* (URL) suele ser también un buen indicador del país.

En el caso de los geoportales locales se ha observado que (3), (4) y (5) son la fuente de los nombres geográficos que deben ser tratados. El nombre geográfico de mayor frecuencia dentro del conjunto compuesto por (3), (4) y (5) representa una región de un país (administrativo y/o geográfico) el cual describe la extensión geográfica del geoportal. Los nombres de países son escasos, y si aparecen, uno de ellos suele contener el identificador de región.

Ámbito	URL	E.G. Manual	(1) N.G. Geo Meta	(2) N.G. `META`	(5) N.G. `BODY`	(3) N.G. `ANCHOR`	(4) N.G. `IMG alt`
Global	http://www.unrisd.org	World	-	UK United Nations UN	UN Brazil (x2) France Yonsei University Seoul Geneva, Switzerland Paris	Brazil	-
Global	http://www.natureserve.org	World	-	-	U.S. Canada Latin America America (x4) Portland,	America	America Colorado

					Oregon		
					Washington, D.C.		
Global	http://www.grid.unep.ch/	World		United Nations	Geneva (x4)	Europe (x4)	
					Europe (x2)	Switzerland	
					Kazakhstan		
					Paris		
					France		
					Switzerland		
					Istanbul		
					Turkey		
					Rabat		
					Morocco		
					New Delhi		
					India		
Regional	http://www.uneca.org/disd/ict/index.htm	Africa	-	-	Africa (x23)	Africa (x7)	-
					Ethiopia (x2)	Ghana	
					Addis Ababa (x8)	Nigeria	
					Nairobi, Kenya	Ethiopia	
					Ghana (x4)	Swaziland.	
					UN (x5)	University of Copenhagen	
					Kingdom of Swaziland		
					Southern Africa		
					University of Copenhagen		
					Luanda, Angola		
Regional	http://www.eurogi.org/	Europe			Europe (x2)	Serbia	
					Bratislava	Beograd	
					Slovakia	Abu Dhabi UAE	
						Ostrava,Czech Republic	
Nacional	http://www.nls.fi/ptk/infrastructure/	Finland	-	-	-	Helsinki	-
Nacional	http://www.fgdc.gov/	USA	-	-	U.S.	North American	USA
Nacional	http://www.igm.cl/	Chile	-	-	Santiago Centro (x2)	Chile (x2)	.
Nacional	http://snig.igeo.pt/portal/	Portugal	-	-	Portugal		
Local	http://www.isgs.uiuc.edu/nsdihome/	Illinois			Illinois (x5)	Chicago (x2)	Illinois
						Illinois (x13)	Urbana-Champaign

Local	http://www.geoportal-idec.net/geoportal/cas/	Cataluña		-	Cataluña Catalunya (x2) Parc de Montjuïc Barcelona (x2)	Europa	Cataluña Catalunya
Local	http://www.ntlis.nt.gov.au	Northern Territory			Northern Territory Australia	Northern Territory	Northern Territory
Local	http://www.iderioja.larioja.org/	La Rioja	La Rioja Logroño ES-LO 42.27189379;-2.28201889	La Rioja (x6) Rioja (x2)	La Rioja (x5) Valencia España Unión Europea	España La Rioja (x2)	La Rioja (x2)
Local	http://www2.idepa.es/idesipa/	Asturias	-	Asturias	UN		EU

Tabla 16: Análisis para el desarrollo de heurísticas para estimación de la Extensión Geográfica

(E. G.: Extensión Geográfica N. G.: Nombre Geográfico)

Anexo V.II. Heurísticas

Normalización de Nombres Geográficos

En primer lugar, antes de aplicar cualquier algoritmo para la estimación del *Bounding Box* es necesario normalizar el conjunto de los nombres geográficos obtenidos como resultado en cada uno de los extractores considerados.

Para la extracción del elemento de *Bounding Box* se ha desarrollado y utilizado un componente de codificación geográfica simple basado en un *geocoder* externo: *Geocoder Compuesto* [57]. El *Google Geocoder* (API de Google *Maps*[47]) ha sido seleccionado como conector principal del *Geocoder Compuesto* por sus características:

- proporciona cobertura de todo el mundo,
- su modelo proporciona organización territorial,
- da preferencia a las más importantes y
- devuelve los resultados en un idioma (presentación de resultados de forma contextual).

Por lo tanto, la herramienta desarrollada utiliza el *Geocoder Compuesto* como fuente de información sobre la organización territorial y también permite solventar la problemática de la ambigüedad de topónimos, dar soporte plurilingüe y tratar algunos errores de la herramienta NER. El *geocoder*, por defecto, genera por cada búsqueda un listado de Entidades Geográficas (EG) ordenado según la preferencia de Google. Por requerimientos del proyecto, la configuración sólo toma en cuenta la primera EG del listado obtenido, una vez filtrado por los siguientes tipos asignados por Google: `COUNTRY`, `REGION`, `SUB_REGION` y `TOWN`. Por cada nombre geográfico se obtiene como resultado una EG normalizada siguiendo el modelo de Google, a la cual se le añade la región global (una o varias) según el código del país desde una Lista de Regiones Globales. La estructura resultante de la entidad geográfica se detalla a continuación:

[*alias de la entidad geográfica*]
[:*tipo asignado*]
[ORG:*nombre original de la búsqueda*]
[REG_GLOB:*región global1,región global2,…*]
[COUNTRY:*código ISO del país*]
[REGION:*región administrativa*]
[SUB_REGION:*subregión administrativa*]
[TOWN:*localidad*]

Por ejemplo:

```
[Zaragoza]
[:TOWN]
[ORG:ZARAGOZA]
[REG_GLOBAL:Europe,EU,UN]
[COUNTRY:ES]
[REGION:Aragón]
[SUB_REGION:Zaragoza]
[TOWN:Zaragoza]
```

[47] http://code.google.com/intl/es-ES/apis/maps/documentation/geocoding/

La Lista de Regiones Globales, utilizada en el proceso de normalización de los nombres geográficos, incluye un listado controlado de todos los continentes y organismos territoriales internacionales de mayor interés. Para cada uno de ellos su estructura contempla la siguiente información:

- países (código ISO) que conforman la región,
- denominación en otros idiomas y
- *Bounding Box* de la región.

[H1] Heurística General:

> *Si se encuentran metadatos geográficos (1), sus valores son prioritarios dando la preferencia a los metadatos de posición (latitud/longitud). En caso de su ausencia, se utilizan los nombres geográficos encontrados (con mayor frecuencia) en los elementos* `META` *no geográficos (2) para generar la extensión geográfica (Bounding Box).*

A continuación se presenta el algoritmo (en pseudocódigo) que detalla la heurística H1.

```
INPUT: MG[..] // Conjunto de GeoMetas del Extractor (1)

Si card(MG[..])>0 Entonces
      Pos[..] = GetPos(MG[..]) // Conjunto GeoMetas de Posición
      Si card(Pos[..])>0 Entonces
            BBOX[..] = BBOX(Pos[..])
            Code = "Estimado"
      Si No Entonces
            EG[..] = MG[..]
            BBOX[..]= getBBOX(EG[..])
            Code = "Estimado"
      Fin Si
Si No Entonces

      INPUT: EG2[..] // Conjunto de EG del Extractor (2)

      EG[..] = getMaxFreq(EG2[..])
      Si card(EG[..])>0 Entonces
            BBOX[..] = getBBOX(NG1[..])
            Code = "Estimado"
      Si No Entonces
            Si card(EG2[..]>0 Entonces:
                  BBOX[..]= getBBOX("World")
                  Code = "Asignado"
            Fin Si
      Fin Si
Fin Si
OUTPUT: BBOX[..] // Conjunto de BBOX
        Code // Código de procesamiento
        EG[..] // Conjunto de EG final (Para su evaluación)
       // Fin
```

Funciones:

`card(X[..])`	Calcula la cardinalidad del conjunto X
`getMaxFreq(X[..])`	Calcula subconjunto de EG con mayor frecuencia del conjunto X
`getBBOX(X[..])`	Obtiene el *Bounding Box* para cada EG del conjunto X

[H2] Heurística Simple:

La extensión geográfica de la página Web se estima a base del nombre geográfico de mayor frecuencia. En caso de que no exista un único nombre geográfico se agrupan los nombres geográficos según la organización territorial a la cual pertenecen empezando desde el menor nivel hacia arriba.

A continuación se presenta el algoritmo (en pseudocódigo) que detalla la heurística H2.

```
INPUT: EG3[..] // Conjunto de EG del Extractor (3)
       EG4[..] // Conjunto de EG del Extractor (4)
       EG5[..] // Conjunto de EG del Extractor (5)
Si card(sum(EG3[..],EG4[..],EG5[..]))>0 Entonces
      // Se calcula la frecuencia principal
      EG[..]= getMaxFreq(Sum(EG3[..],EG4[..],EG5[..]))
      Si card(EG[..])=1 Entonces
            BBOX[..] = getBBOX(EG[..])
            Code = "Estimado"
      Si No
            // Hay que reagrupar para calcular frecuencia
            Max = GetMaxLevel(EG[..])
            Code = "Asignado"
            Level = GetMinLevel(EG[..])+1
            Mientras (Level<=Max AND Code=="Asignado")  Hacer
                  EG[..] = getMaxFreq(GetLevel(EG[..],Level))
                  Si card(EG[..])=1 Entonces
                        BBOX[..] = getBBOX(EG[..])
                        Code = "Estimado"
                  Si No
                        Level = Level + 1
            Fin Mientras
            // Si todavía no se ha estimado, se asigna.
            Si (Code == "Asignado")
                  BBOX[..]= getBBOX("World")
            Fin Si
Si No
      BBOX[..]= getBBOX("World")
      Code = "Asignado"
Fin Si
OUTPUT: BBOX[..] //Conjunto de BBOX
        Code // Código de procesamiento
        EG[..] // Conjunto de EG final (Para su evaluación)
// Fin
```

Funciones:

`sum(X[..],Y[..],…)`	Suma de los conjuntos X, Y, …
`card(X[..])`	Calcula cardinalidad del conjunto X
`getMaxFreq(X[..])`	Calcula subconjunto de EG con mayor frecuencia del conjunto X
`getLevel(X[..],L)`	Retorna subconjunto de nivel L de organización del conjunto X
`getMinLevel(X[..])`	Nivel de organización territorial mínimo del conjunto X
`getMaxLevel(X[..])`	Nivel de organización territorial máximo del conjunto X
`getBBOX(X[..])`	Obtiene el *Bounding Box* para cada EG del conjunto X

Nota: En caso de usar H1 y H2, si H1 no asigna BBOX (`Code=="Estimado"`), se ejecuta H2.